SELECTED PAPERS ON RESPONSE SURFACE METHODOLOGY

Robert C. Williges, et al

Illinois University

Prepared for:

Air Force Office of Scientific Research

August 1972

Technical Report ARL-72-20/AFOSR-72-9

August 1972

Research supported by:

Air Force Office of Scientific Research
Contract F44620-70-C-0105

SELECTED PAPERS ON RESPONSE SURFACE METHODOLOGY

Edited by

Robert C. Williges
Beverly H. Williges

Prepared for:

Life Sciences Program
Air Force Office of Scientific Research

University of Illinois at Urbana-Champaign
AVIATION RESEARCH LABORATORY
INSTITUTE OF AVIATION
University of Illinois-Willard Airport
Savoy, Illinois

151

UNCLASSIFIED

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Aviation Research Laboratory, Institute of Aviation
University of Illinois
Urbana, Illinois 61801

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP

3. REPORT TITLE
SELECTED PAPERS ON RESPONSE SURFACE METHODOLOGY

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Scientific Interim

5. AUTHOR(S) (First name, middle initial, last name)
Robert C. Williges and Beverly H. Williges (Editors)

6. REPORT DATE
August 1972

7a. TOTAL NO. OF PAGES
~~139~~ 151

7b. NO. OF REFS
70

8a. CONTRACT OR GRANT NO. F44620-70-C-0105

b. PROJECT NO. 9778

c. 61102F

d. 681313

9a. ORIGINATOR'S REPORT NUMBER(S)

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
AFOSR - TR - 72 - 2441

10. DISTRIBUTION STATEMENT
Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES
TECH, OTHER

12. SPONSORING MILITARY ACTIVITY
Air Force Office of Scientific Research
1400 Wilson Boulevard (NL)
Arlington, Virginia 22209

13. ABSTRACT

Ia

UNCLASSIFIED

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Response Surface Methodology						
Central-Composite Design						
Experimental Design						
Multiple Regression						
Transfer of Training						
Pursuit Rotor						
Epicycloid Pursuit Rotor						
Distribution of Practice						
Degree of Original Learning						
Task Difficulty						
Television Displays						
Television Focus						
Nontarget Density						
Television Raster Lines						
Visual Angle						
Image Interpretability						
Cartographic Symbology						
Surveillance System						
Visual Time Compression						
Radar						
Target Introduction Rate						
Radar Clutter Density						
Radar Target Velocity						
Radar Blip/Scan Ratio						
Radar Clutter Replacement Probability						

SELECTED PAPERS ON RESPONSE SURFACE METHODOLOGY

Edited by

Robert C. Williges
Beverly H. Williges

Approved for public release; distribution unlimited.

I c

FOREWORD

The five papers included in this technical report constitute the original manuscripts submitted to Human Fuctors for regular journal publication. Hopefully, any inconsistencies and errors that may be present will be corrected before any of the articles appear in print.

Although each paper was purposefully written as a complete independent paper, all of the papers taken together summarize much of the research effort to date on one task of a current contract with the Air Force Office of Scientific Research. This project is one of eight tasks in a contract titled "The Enhancement of Human Effectiveness in System Design, Training, and Operation." Four of the tasks are in the area of pilot selection, training, and performance assessment, and four deal with avionics system design principles.

The papers have been arranged in this report to show the sequence of the research effort. The first manuscript, Clark and Williges (1972), is an introductory paper. Based on an article published by Williges and Simon (1971), the purpose of the Clark and Williges (1972) paper is to introduce the Response Surface Methodology (RSM) central-composite design and to consider various design modifications necessary for using RSM central-composite designs in human performance research. The remaining four papers both illustrate the use of RSM central-composite designs for developing multiple regression prediction equations and empirically test some of the design modifications suggested by Clark and Williges (1972).

The Williges and Baron (1972) manuscript reports a between-subjects, RSM central-composite design for human transfer of training assessment and demonstrates the advantage of replicating the design across all data points. Reporting a within-subject, RSM central-composite design, the Williges and North (1972) paper compares collapsed and uncollapsed data analyses in terms of sensitivity and predictive validity as determined through cross-validation.

The last two papers, Mills and Williges (1972) and Williges and Mills (1972), are concerned with research sponsored by the Aerospace Medical Research Laboratory,

Aerospace Medical Division, Air Force Systems Command, Wright-Patterson AFB and appear as AMRL Technical Reports. Additional support for data analyses was provided by the Air Force Office of Scientific Research on the current contract with the Aviation Research Laboratory of the Institute of Aviation, University of Illinois at Urbana-Champaign. The Mills and Williges (1972) paper illustrates a rather complex use of a within-subject, RSM central-composite design to predict performance in a single-operator simulated surveillance system. The last paper, Williges and Mills (1972), evaluates the predictive validity of the multiple regression equations of the previous study in terms of predictive accuracy to other data points within the range of the variables originally tested.

A number of people were quite helpful in the preparation of these papers. Specific acknowledgments to many of them are provided at the end of each manuscript. Five additional people, however, deserve special mention. Dr. Stanley N. Roscoe and Dr. Melvin J. Warrick provided valuable comments on different aspects of some of the papers. Mrs. Tatie Wrobel proofread and made additional editorial comments on all the papers. Mr. Morris Maitland diligently prepared all the final figures. And, Mrs. Carolyn Gardner was able to remain in good spirits after expertly typing and retyping each manuscript a countless number of times.

References

Clark, C. and Williges, R. C. Response surface methodology central-composite design modifications for human performance research. Human Factors, 1972, submitted for publication.

Mills, R. G. and Williges, R. C. Performance prediction in a single-operator simulated surveillance system. Human Factors, 1972, submitted for publication.

Williges, R. C. and Baron, M. L. Transfer assessment using a between-subjects central-composite design. Human Factors, 1972, submitted for publication.

Williges, R. C. and Mills, R. G. Predictive validity of central-composite design regression equations. Human Factors, 1972, submitted for publication.

Williges, R. C. and North, R. A. Prediction and cross-validation of video cartographic symbol location performance. Human Factors, 1972, submitted for publication.

Williges, R. C. and Simon, C. W. Applying response surface methodology to problems of target acquisition. Humnn Factors, 1971, 13, 511-519.

Response Surface Methodology Central-Composite Design Modifications for Human Performance Research

CHRISTINE CLARK and ROBERT C. WILLIGES, University of Illinois at Urbana-Champaign

Selected Response Surface Methodology (RSM) designs that are viable alternatives in human performance research are discussed. Two major RSM designs that are variations of the basic, blocked, central-composite design have been selected for consideration: 1) central-composite designs with multiple observations at only the center point, 2) central-composite designs with multiple observations at each experimental point. Designs of the latter type are further categorized as: a) designs which collapse data across all observations at the same experimental point; b) between-subjects designs in which no subject is observed more than once, and observations at each experimental point may be multiple and unequal or multiple and equal; and c) within-subject designs in which each subject is observed only once at each experimental point. The ramifications of these designs are discussed in terms of various criteria such as rotatability, orthogonal blocking, and estimates of error.

INTRODUCTION

Frequently, an investigator's aim is to determine a quantitative relationship between human performance and one or more system parameters. Among the most immediate benefits accruing from such a known, quantitative relationship are the ability to predict performance levels corresponding to given levels of the system variables and, conversely, the ability to determine the system variable levels necessary to maintain a designated performance level. One particularly promising procedure for gathering the data needed to make these and other quantitative determinations is Response Surface Methodology (RSM), originally introduced by Box and Wilson (1951). Unlike traditional factorial analysis of variance designs, RSM focuses primarily on determining the functional relationship that exists between the response and specified continuous, quantitative factors, rather than merely determining the significance of the various factors.

In addition to approximating the relationship between performance and factors in the form of a prediction equation, RSM advances a variety of experimental designs to achieve that estimate as efficiently and economically as possible. When using factorial designs, the investigator is often forced by practical considerations to limit the number of factors studied to even less than the number that he believes has a critical effect on performance. In such a case he must conduct multiple studies, each of which investigates only a few factors at any one time. This results in an unrealistic view of any system in which factors are not independent of one another. By allowing the investigator to consider larger numbers of factors within a single study, RSM proves a valuable investigatory tool. Through strategic sampling of data points, RSM also provides the most essential information and allows one to decide whether or not the collection of additional data is merited.

Most RSM designs are special cases of the Box and Wilson (1951) central-composite design. Although this design was originally developed for application in chemical research, its utility in psychological research, especially in studies of human performance, has been documented (Meyer, 1963; Simon, 1970; Williges and Simon, 1971). It is not unreasonable, however, to anticipate the need for some modification in that basic design to make it more appropriate for research involving human subjects. The purpose of this paper is to suggest several appropriate design modifications that attempt to retain as many of the positive traits of the RSM central-composite design as possible. Before discussing these modifications, a description of central-composite designs is necessary.

CENTRAL-COMPOSITE DESIGNS

Suppose that an investigator were interested in predicting radar target detection, Y, given level s of display resolution, X_1, visual angle, X_2, and random noise, X_3. Further suppose that the true relationship between target detection and the three display-related variables could be expressed as a function f of the levels of X_1, X_2, and X_3. That is, in symbolic form

$$Y = f(X_1, X_2, \ldots, X_m) + e,$$

where $m = 3$; X_i, $i = 1, 2, 3$, is the level of the i^{th} display-related variable; e is the associated experimental error; and Y is the corresponding level of target detection. The particular function which describes the relationship in question is called the <u>response surface</u>. Of course, in practice one usually does not know just what that function is. Therefore, the investigator attempts to derive a reasonabie estimate of the unknown function, basing his estimate upon the examination of representative data. In other words, the investigator attempts to approximate the response surface, the true functional relationship between response and factor levels, by using a derived polynomial equation. For example, in lieu of the function f, he might substitute a complete second-order polynomial in X_1, X_2, and X_3 of the form

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_1^2 + b_5X_2^2 + b_6X_3^2 + b_7X_1X_2 + b_8X_1X_3 + b_9X_2X_3 ,$$

where the numerical values of b_0 through b_9 are determined empirically according to multiple regression techniques. The complete second-order ploynomial includes the linear effect of each variable, the linear by linear interactions, and the quadratic effect of each variable.

Factorial Design: A Data Collection Procedure

When developing an equation to approximate the response surface, the investigator measures the desired response at relatively few data points, each designated by some unique combination of independent variable or *factor* levels. For example, the investigator studying target detection might adopt a factorial design in which each of the three display-related variables assumes two levels, -1 and +1. Of course, these two factor levels can represent any desired real-world factor levels simply by applying the appropriate linear transformation. Determination of real-world factor levels using such a transformation is illustrated in a later section. The 2^3, or 8, possible combinations of factor levels designate the particular set of points at which the investigator measures the response. In simple terms, the factorial design serves as a set of directions for collecting data.

If the factors are continuous and quantitative, the data collected in this manner can serve as the raw input data for either a traditional analysis of variance or a multiple regression analysis. When the investigator's aim is to derive a polynomial approximation to a response surface, rather than merely to determine the significance of the various factors, multiple regression is the more appropriate analysis. The factorial design provides the quantitative levels of the relevant factors or *predictor variables*, and the investigator makes direct measurements of the response level at each data point designated by the design. In the case of the preceding example, because each of the three factors, display resolution, visual angle, and random noise, assumes two distinct, quantitative levels, a first-order polynomial equation in each factor can be fitted to the data.

If the investigator suspects that target detection is at least a complete second-order function of the three display-related factors, he must measure detection performance at more than two levels of each of those variables. He could, for example, provide for a complete second-order equation in all three factors by collecting the appropriate data according to another factorial design in which each factor assumes three levels. Such a design designates a total of 3^3 or 27 points at which target detection performance is measured, an increase of 19 data points over the previous design.

Central-Composite Design: An Alternative Data Collection Procedure

An alternative procedure could be followed to direct data collection efforts. Suppose the investigator maintained the initial two-level factorial design involving only eight unique factor combinations. He could augment that basic design by including the following $(2 \cdot 3 + 1)$ or 7 additional distinct factor combinations, expressed here as ordered triplets of factor levels:

$(0, 0, 0)$;
$(-\alpha, 0, 0)$; $(\alpha, 0, 0)$;
$(0, -\alpha, 0)$; $(0, \alpha, 0)$;
$(0, 0, -\alpha)$; and $(0, 0, \alpha)$.

Again, these factor levels can represent any desired real-world factor levels simply by applying the appropriate linear transformation. The numerical value which α assumes is chosen so as to insure certain advantageous design properties to be discussed later. The particular α value is not crucial to the current discussion; suffice it to say at this point that α is merely one of the levels which the factors can assume.

The addition of these seven new data points to the basic factorial design results in a design composed of 15 distinct factor combinations. Yet the investigator can now fit not only a second-order polynomial to the resulting data, but also a polynomial involving some higher-order predictors as well. This is usually more than adequate for approximating most response surfaces. With an increase in only seven in the number of distinct data collection points the investigator is able to measure the response at five levels of each factor, those five levels being the values $\pm\alpha$, ± 1, and 0. (The corresponding complete factorial design involving five levels of each factor entails 125 distinct data points for a single replication.) Moreover, if repeated observations were made at the center point (0, 0, 0), the resulting design would provide for an estimate of experimental error variance. This error estimate allows the investigator to test the significance of the derived polynomial and each of its components, as well as testing the significance of effects not included in the derived equation.

This proposed alternative design is merely a combination or composite of a traditional 2^3 factorial design and some strategically selected additional points (Box and Wilson, 1951). In particular, the design is a three-factor central-composite design in that the designated factor combinations or data points are spaced symmetrically about a central or center point designated by the ordered triplet of factor levels (0, 0, 0) as shown in Figure 1. More generally, a K-factor central-composite design is realized by combining a basic 2^K factorial with the $(2 \cdot K + 1)$ additional distinct factor combinations

$$
\begin{array}{ll}
(0, 0, \ldots, 0); & (-\alpha, 0, \ldots, 0); \; (\alpha, 0, \ldots, 0); \\
 & (0, -\alpha, \ldots, 0); \; (0, \alpha, \ldots, 0); \\
\ldots & (0, 0, \ldots, -\alpha); \; (0, 0, \ldots, \alpha)
\end{array}
$$

(Cochran and Cox, 1957, p. 343).

Note that each of the 2K noncenter points is defined such that all factors except one are held at the 0 level, whereas the remaining factor assumes the values $-\alpha$ and $+\alpha$, in turn. The aggregate of these 2K additional noncenter points is referred to as the star or axial portion of the resulting central-composite design. As the number of factors increases to five or more, a $2^{(K-p)}$ fractional factorial, where p is a positive integer, is often substituted for the complete 2^K factorial, thereby reducing still further the number of distinct data points (see Cochran and Cox, 1957, Ch. 6A). In such instances, a K-factor central-composite design is realized by combining a $2^{(K-p)}$ fractional factorial with the same $(2 \cdot K + 1)$

combinations given above. More specifically, when fractional factorials are incorporated into a second-order central-composite design, one chooses the defining contrast such that all the first- and second-order components are present and are not aliases of each other. Were this restriction not observed, the first- and second-order effects would be inextricably mixed with one another. Regardless of the number of factors, however, each factor assumes five distinct levels corresponding to the coded values $\pm\alpha$, ± 1, and 0. Moreover, the designated factor combinations fall symmetrically about the center point (0, 0, ..., 0).

- - - - - - - - - - - - - - - - - -

Insert Figure 1 about here.

- - - - - - - - - - - - - - - - - -

Again, if the factors and the response are continuous, quantitative entities, the data can be analyzed using multiple regression techniques. To test for the significance of the derived polynomial and its components and the significance of all other terms not included in the equation, the investigator needs an estimate of experimental error variance. The central-composite design provides for an estimate of error by repeating observations at the center point (0, 0, ..., 0). Choosing the appropriate number of replications results in a design in which the standard error of estimate is roughly the same at all points within the experimental region. Hence, the estimate of error at the center is used as an estimate of error throughout the entire K-space, thereby minimizing redundancy. Too many replications at the center yield standard errors of estimate which increase rapidly for those points farther from the center. On the other hand, with too few replications of the center point, the standard error is apt to be greater at the center than at the surrounding data points. In the case of a three-factor central-composite design, for example, the suggested number of replications at the center point is six, thereby increasing the total number of observations to 20. See Table 1. Although the derivation procedures are beyond the scope of this discussion, procedures exist for determining the optimum number of center points of a K-factor design (Box and Hunter, 1957).

- - - - - - - - - - - - - - - - -

Insert Table 1 about here.

- - - - - - - - - - - - - - - - -

Design Limitations

Of course, reducing the size of an experiment by eliminating data points has its price. Coincident with the reduction in data is a reduction in obtained information. In particular, when fractional factorials are incorporated into the central-composite design, at least one factorial effect, the defining contrast, is lost entirely. Prudent choice of the defining contrast(s), however, usually results in losing information concerning some higher-order interaction(s) which seldom affect performance anyway. In addition, interpretation of that information which is provided by a fractional factorial central-composite design is somewhat more ambiguous in that certain effects are mixed with one another, as indicated above. By choosing the highest-order interaction as the defining contrast, the experimenter can insure that first- and second-order effects are not confounded with one another.

Rotatability

One desirable property of some central-composite designs is rotatability (Box and Hunter, 1957). Rotatability exists when there is equal reliability of predicted responses at all data points equidistant from the center. This is an especially convenient design quality in exploratory work when the investigator is ignorant of the response surface and its relative orientation to the orthogonal factor axes. Rotatability imposes the additional constraint on factor level selection that the value of α be equal to $2^{K/4}$ (Box and Hunter, 1957). When a $2^{(K-p)}$ fractional factorial design is used in place of the full 2^K factorial, then α must equal $2^{(K-p)/4}$ if rotatability is to exist (Box and Hunter, 1957). Thus, if the hypothetical three-factor design diagrammed in Figure 1 is to be rotatable, the α value must be 1.682, because $2^{K/4} = 2^{3/4} = 8^{1/4} = 1.682$. To insure roughly equal precision of prediction across the entire experimental region, the center point is replicated six times. When complete, the design involves a total of 20 observations (as indicated in Table 1) with 14 of the experimental factor combinations lying on the surface of a sphere of radius 1.682, and with 6 observations being made at the center point (0, 0, 0).

Selection of Factor Levels

The first, and perhaps most crucial, step in selecting factor levels for a central-composite design (or even a basic factorial) is to determine the experimental range of each factor to be incorporated into the design. Because polynomials cannot be extrapolated with confidence, the derived polynomial equation should be considered an approximation to the response surface only within the region defined by the respective factor ranges. When appropriately transformed, the limiting real-world values of each factor, as determined by the selected range, yield the coded values $-\alpha$ and $+\alpha$, and the center of that range yields the coded value 0. For example, suppose that the values of interest for display resolution range from 168 to 504 TV lines/dm. Further suppose that $\pm\alpha$ assume the values -1.68 and +1.68 respectively, so as to insure that the resulting design is rotatable. The investigator's next task is to determine the linear transformation which: (a) when applied to the center of the factor range, 336, yields the coded value 0, and (b) when applied to the lower and upper limiting values of display resolution, 168 and 504, yields the coded values -1.68 and +1.68, respectively. It can be demonstrated that the following linear transformation satisfies both these requirements:

$$X_1^* = \frac{X_1 - 336}{100} ,$$

where X_1^* is a coded factor level and X_1 is the corresponding real world factor level. The remaining two levels of display resolution are determined by solving for X_1 where X_1^* assumes the values -1 and +1 in turn. Therefore, the appropriate five real-world levels of display resolution are 168, 236, 336, 436, and 504 TV lines/dm.

The appropriate real-world levels of all other experimental factors are determined in like manner. In each case, (a) the range of the factor and the center point are established, (b) the appropriate linear transformation is determined, and (c) the remaining two levels of the factor are determined in accordance with the transformation. Although coding the appropriate real-world factor levels once they are determined is not necessary, the use of linear transformations of the data simplifies analysis without affecting the result of any subsequent statistical tests. On occasion this rigid demand regarding the selection of data points makes the central-composite design impractical for some human factors studies. For example, variables such as target type, target complexity, and briefing instructions are not readily quantifiable. Moreover, it is sometimes neither practical nor feasible to measure even certain quantifiable variables at the five levels specified by the central-composite design. Alternative RSM designs have been developed which require fewer than five levels (Box and Behken, 1960, and Draper and Stoneman, 1968).

Blocking

An additional feature of central-composite designs that affords the investigator greater efficiency and flexibility is blocking. Under blocking conditions, subsets of the complete set of data collection points are studied together. If the blocking is orthogonal, any differences in mean performance among blocks are independent of any main effects due to the independent variable manipulations, and as such, they do not affect the underlying quantitative relationship between factors and performance. If blocking were not orthogonal, the derived prediction equation would be a function of block effects as well as main effects. This aspect of design is valuable to the human factors engineer who is concerned with isolating potential effects due to such factors as different experimenters, changes in apparatus, and variable environmental conditions. Recall the investigator studying radar target detection as affected by display resolution, visual angle, and random noise. It is unlikely that all the necessary data can be collected during a single flight or perhaps not even in the same aircraft. By taking advantage of orthogonal blocking techniques, he can guard against the parameters of the derived prediction equation being affected by such differences. For example, a block could refer to that set of observations which were made during any given flight.

Blocking of a central-composite design is accomplished readily by subdividing the design into two parts: (a) the 2^K factorial (or $2^{(K-p)}$ fractional factorial) portion and (b) the set of 2K points comprising the star or axial portion of the design. As the number of factors increases, the 2^K factorial (or $2^{(K-p)}$ fractional factorial) can be subdivided further into additional blocks by using fractional factorials. When fractional factorials are used for blocking second-order designs, care must be taken not to confound any first- or second-order effects with blocks, and none of these effects should be aliases of one another within a given block.

Orthogonal blocking placed additional constraints on the central-composite design concerning the selection of α and the number of center points. These parameters must be chosen to insure that the average predicted response level is the same for every block. Orthogonal blocking is guaranteed when the following condition is met (Box and Hunter, 1957, p. 230):

$$\frac{2\alpha^2}{2^K} = \frac{(N_s + N_{s0})}{(N_c + N_c 0)} , \qquad (1)$$

or, in the event that a $2^{(K-p)}$ fractional factorial is incorporated into the design,

$$\frac{2\alpha^2}{2^{(K-p)}} = \frac{(N_s + N_{s0})}{(N_c + N_{c0})} , \qquad (2)$$

where N_{c0} and N_{s0} are the number of center points added to the intact 2^K factorial portion and the 2K star portion of the design, respectively. N_c and N_s reflect the number of noncenter points in the 2^K factorial and in the 2K star, respectively.

Given the proposed design in Figure 1 for studying radar target detection, orthogonal blocking can be achieved by dividing the 20 data points given in Table 1 into subsets of 6, 6, and 8 observations, as depicted in Figure 2. The first two blocks each represent one-half replicates of the complete 2^3 factorial portion, and the third block is the six-point star portion. Two center points have been included in each of the three blocks for replication. Solving Equation 1 for α yields an α value of 1.633.

- - - - - - - - - - - - - - - - -

Insert Figure 2 about here.

- - - - - - - - - - - - - - - - -

Given this revised value of α, the investigator must revise his choices of real-world factor levels for display resolution. By transforming X_1 where X_1^* assumes the revised α values -1.633 and +1.633, in turn, yields revised levels for the lower and upper limiting values of display resolution; the revised real-world levels are 173 and 499 TV lines/dm, respectively. Similarly, it can be shown that the change in α value does not necessitate a change in the three intermediate real-world values of display resolution. Hence, the five levels appropriate to the orthogonally blocked design are 173, 236, 336, 436, and 499 TV lines/dm.

The investigator must also recompute the appropriate real-world levels of visual angle and random noise in like manner. Note that the value of α required to insure orthogonality is slightly different from the 1.682 value required for rotatability. To achieve orthogonal blocking it is often necessary to sacrifice rotatability, although the appropriate α values are usually quite similar. In human factors applications, however, the potential gains from orthogonal blocking probably outweigh the risk of forfeiting rotatability.

Added flexibility can accrue from use of blocking techniques, as Box and Hunter (1957) illustrated when they employed blocking to facilitate exploration of a response surface. A properly blocked design permits research to be conducted in stages. Each block of data points from the complete second-order design constitutes a first-order, rotatable central-composite design. Gathering data from the first series of blocks, the investigator can judge, for example, whether or not any of the original experimental variables merits being dropped from further consideration or whether or not greater than a linear polynomial is needed to explain the data adequately. If so, the design can be altered here rather than after all data are collected. The ability to make such decisions at an early stage may mean that the investigator is able to conclude his study after collection of considerably less data than he had anticipated.

Analyses

Basically, two standard statistical analyses are conducted on the data accrued from an RSM design. Frist, a least squares multiple regression analysis is performed on the data to determine the functional relationship between performance (Y) and the system variables (X). Multiple regression is merely an extension of simple linear regression such that the multiple regression analysis includes more than one predictor and/or terms other than linear components. Because of the numerical complexity involved in multiple regression, matrix algebra ordinarily is used for the calculation of the regression equation coefficients. In addition, a matrix algebra solution using correlation matrices rather than raw scores provides a flexible and efficient means for handling a variety of possible regression equations within the same computer program. A correlation matrix solution results in a standard regression equation (variables are stated in terms of z scores and the intercept is 0) that can be converted easily into a nonstandard or raw score regression equation.

The second analysis usually performed on data obtained from a RSM design is an analysis of variance performed on the regression analysis. Essentially, the analysis of variance partitions the sums of squares into variation due to regression and variation not due to regression (residual). The regression sum of squares is subdivided into the variation of the particular partial regression weights resulting from the preceding multiple regression analysis. The residual sum of squares can be further subdivided into block effects, subject effects, lack of fit, and error. The main purposes of this analysis of variance are to test the significance of the given partial regression weights and to test for a significant lack of fit which might indicate additional parameters are necessary in the regression equation. All of the sums of squares are converted to mean squares by dividing by the appropriate degrees of freedom. The resulting F ratios are constructed by using the error mean square as the denominator.

Consider again the study of radar target detection, Y, as a function of display resolution, visual angle, and random noise, X_1, X_2, and X_3, respectively. Hypothetical data for such a study are presented in Table 2. A multiple regression analysis of these hypothetical data yields the following generalized, first-order prediction equation:

$$Y = 16.115 - 1.203\, X_1 - 0.503\, X_2 + 0.847\, X_3.$$

Substituting given levels of the independent variables into this equation affords the investigator a corresponding predicted level of detection latency.

- - - - - - - - - - - - - - - - -

Insert Table 2 about here.

- - - - - - - - - - - - - - - - -

The results of a subsequent ANOVA performed on the regression analysis appear in Table 3. The derived equation accounts for nearly 74% of the total variance in detection latency. Each of the coefficients, excluding the constant term b_0, is significant at well beyond the .01 level. Blocks are significant. However, because blocking is orthogonal, the values of the regression weights have not been affected. Noting that the lack-of-fit term is significant, the investigator will submit his data to a second multiple regression analysis to determine a higher-order prediction equation.

- - - - - - - - - - - - - - - - -

Insert Table 3 about here.

- - - - - - - - - - - - - - - - -

For a detailed discussion of the analysis procedures, see Clark and Williges, 1972.

DESIGN CONSIDERATIONS

In a recent article, Williges and Simon (1971) discussed several general advantages of the RSM technique which contribute to its potential value in human factors research. Among the most obvious benefits is the economy of data collection. Not only is sampling restricted to the experimental region of greatest interest, but also repeated observations are restricted to the center point of that region. As originally conceived, RSM was developed as a methodology for quickly locating optimums by means of a series of experiments each dependent on the results of the preceding one. More specifically, Box and Wilson (1951) were interested in determining the optimum combination of factor levels needed to produce the maximum yield from a chemical reaction. However, human factors engineers are largely interested in deriving global prediction equations which allow them to predict performance levels accurately throughout an entire range of factor levels.

When the goal is to approximate an entire response surface, rather than merely that portion of the surface surrounding the optimum, limiting multiple observations to a single experimental point may not be the most judicious strategy. Indeed, the actual variability in response may be so great across subjects and data points, that to presume the standard error of estimate at the center point as an adequate estimate of error at all points is unrealistic. A recent study concerning transfer of training (Williges and Baron, 1972) affords a striking demonstration of the effect of estimating experimental error at a single replicated point as opposed to estimating it across a series of replicated points. When replications were restricted to the center point, none of the experimental factors was found to contribute significantly to the response level, despite their apparent importance in the resulting prediction equation. When multiple observations were made at each of the data points, however, the subsequent analysis revealed that some of the experimental variables were significant in determining the response level. Of course, when the basic RSM central-composite design is modified in such a manner, methodological questions arise concerning how best to retain the positive attributes of the basic design, while still making the modifications appropriate to research with human subjects. For example, should repeated observations be made at more than one experimental point; should all data be retained or should they be collapsed; should different subjects be observed at each experimental point or should the same subjects be observed at all points; under what conditions are particular design variations especially appropriate?

The following discussion proposes several design variations appropriate to human factors research together with the ensuing methodological considerations. A generalized computer program to analyze data from each of these design variations as well as data from the basic RSM central-composite design has been developed by Clark, Williges, and Carmer (1971), and a detailed discussion of the statistical procedures is presented by Clark and Williges (1972).

Collapsed Designs

The simplest modification is achieved merely by replicating the entire central-composite design a given number of times. Consider, for example, the orthogonally blocked, RSM central-composite design depicted in Figure 2. Suppose the investigator elects to replicate that design five times. The data points remain the same as those listed under Figure 2. Now, however, the design involves a total of 100 observations, over a total of 15 distinct factor combinations. Block 1 now contains 30 observations, Block 2 contains 30 observations, and Block 3 contains 40 observations. Note that, although multiple observations have been made at each of the experimental points, the center point has still been replicated six times more than any other point. Although the points on the surface of the sphere have been replicated 5 times, the center point has been replicated 30 times, 10 times within each of the three blocks.

At this point the investigator must decide whether or not to retain and analyze directly the data corresponding to all 100 observations. He could collapse his data across those subjects within the same block who were observed at the same experimental point and then analyze the collapsed data without having to make any modifications in calculation procedures. The net effect of collapsing in this manner is a data matrix identical in form and number of observations to one resulting from the original blocked RSM central-composite design shown in Figure 2. Now, however, the data are combined values obtained from collapsing rather than values representing a simple observation. In addition, estimates of experimental error are obtained from the resulting six center points, each of which is a collapsed score.

This procedure has the advantage of retaining all the features of a RSM central-composite design as well as adding stability to the experimental data points because the collapsed data are not heavily biased by the results of any one extreme subject. This is especially valid if the median is used as the combining statistic. Because it is probably of little value to develop unique prediction equations for each subject, such a collapsing procedure may be appropriate even though degrees of freedom are lost from the design.

A recent cross-validation study (Williges and North, 1972), however, illustrates a potential drawback of collapsing data prior to analysis. When median data were used to derive prediction equations, the resulting multiple regression coefficient R was notably higher than the corresponding value resulting from the comparable noncollapsed data analysis. However, the shrinkage of R from the original sample to the cross-validation sample was very pronounced when regression was based on collapsed data. There was far greater shrinkage than that predicted by the modified Wherry shrinkage formula (Lord and Novick, 1968; Herzberg, 1969). On the other hand, shrinkage was minimal when derivation was based upon noncollapsed data. Hence, for predicting response levels for individuals not included in the derivation sample, the collapsed analysis did not afford appreciably better prediction despite the deceivingly greater accuracy of the derived prediction equation as suggested by the initially high multiple R value. Indeed, the multiple R deriving from noncollapsed data was far more representative of the predictive accuracy of the equation.

Noncollapsed Designs

Suppose that the investigator replicating the blocked central-composite design chooses not to collapse his data across subjects. Rather, he retains each of the subject's data for subsequent analysis. By retaining all this information he gains degrees of freedom for the error term which were previously lost by collapsing the data. Error is now estimated across all points at which replications occur, instead of using only the estimate of the error at the center point as in the collapsed design and the original design. It is quite possible that there may be certain areas of the experimental region in which there is considerable variability in response and other areas in which the variability is negligible. This is particularly true if the range of factor levels under consideration is sizable. Given this variability, it is not reasonable to use the estimate of error at only one area as an estimate of error throughout the experimental region. The prediction equation which one develops should afford a reasonable description of the entire response surface, not merely a selected area of that response surface.

When noncollapsed designs are used, the investigator must make another major decision with respect to his selected design. If, due to the nature of his research problem, he chooses to observe different subjects at each of the experimental points, the resulting study constitutes a between-subjects design. If, on the other hand, he elects to observe each of a set of subjects under all experimental conditions, the resulting study constitutes a within-subject design. The choice of a between- versus a within-subject design is dictated by the particular question which the researcher is investigating. In either case, if the necessary restrictions are observed, the design conforms to the basic central-composite design.

Between-subjects designs. Given certain research questions, observing the same subjects under more than one experimental condition would lead one to draw invalid conclusions concerning the effect of the various experimental manipulations. Consider, for example, an investigation of the comparative efficacy of selected training methods. Certainly Training Method B cannot be evaluated accurately by observing the peformance of subjects who have previously been trained to criterion under Method A, because the observed performance may be a function of not only the condition itself, but also of the preceding condition which he has experienced. In such a case it is imperative that the investigator adopt a between-subjects design, observing each subject under only one experimental condition. The transfer of training study cited earlier (Williges and Baron, 1972) provides such an example.

Recall the detection latency study which replicates the orthogonally blocked central-composite design of Figure 2 five times. If 100 different subjects are observed across those 20 data points (6 of which are the center point), a between-subjects design is realized. Because the full central-composite design is being replicated intact, the necessary relationship guaranteeing orthogonal blocking, as given in Equation 1, is still satisfied. As in the original design the center point is being replicated six times more than any other point. Although experimental error is now being estimated across all data points and includes subject to subject variation, the results of a subsequent analysis to determine a first-order prediction equation are of the same type shown in Table 3. The increased number of observations is reflected in the values for total degrees of freedom, residual degrees of freedom, and error degrees of freedom; the adjusted values are 99, 96, and 83, respectively. Meyer (1963) has used this design procedure successfully in a human learning experiment.

If, indeed, the variability in response at each of a series of data points is used as an estimate of experimental error variance, there is no need to replicate one point more than any other. In the original central-composite design, in which only the center point is replicated, the additional observations at that point provide the investigator with his only estimate of error. But, with repeated observations occurring at each of the experimental points, there appears no need to make more observations at the center merely for the sake of obtaining an estimate of error. The investigator could choose instead to replicate each of the experimental points, including the center, an equal number of times, while still maintaining the use of different subjects for each observation.

Eliminating observations at the center point, however, has implications for orthogonal blocking. It is now necessary to adjust the value of α accordingly, because the original blocking has been disturbed due to the elimination of center points from the factorial portion of the design and the reduction in the number of center points in the star portion of the design. With respect to the target detection latency example in which repeated observations are made at each of 15 unique experimental points, making the appropriate adjustment results in an α value of 1.87 rather than 1.633, as defined by Equation 1. This change in the α value is reflected in Figure 3 which designates the orthogonal blocking of the 15 unique experimental points. Note the reduction of data collection points within each of the three blocks, and the complete absence of center points in Blocks 1 and 2.

Changing the coded value of α also necessitates reselecting the real-world levels of the various factors under study. Recalculating the levels of display resolution, for example, the investigator learns that the five levels appropriate to the new orthogonally blocked design are 149, 236, 336, 436, and 523. Selecting these five levels retains the center of the experimental region, but increases its range beyond that indicated in Figure 2.

- - - - - - - - - - - - - - - - -

Insert Figure 3 about here.

- - - - - - - - - - - - - - - - -

Replicating this modified RSM central-composite design five times, the investigator makes a total of 75 observations, 20 in Block 1, 20 in Block 2, and 35 in Block 3. Submitting these 75 observations to direct analysis to determine a first-order prediction equation yields results similar to those shown in Table 3. Again, the change in design is reflected in corresponding changes in values of total degrees of freedom, residual degrees of freedom, and error degrees of freedom; the adjusted values are 74, 71, and 60, respectively.

Within-subject design. On occasion the objectives of an experiment make it appropriate and desirable to observe each subject in each treatment condition. In such a case, each individual serves as his own control, and between-subjects variability does not affect the experimental conditions. Moreover, observing the same set of subjects under each treatment condition affords another obvious advantage over the between-subjects designs in that fewer subjects are needed to conduct the study, albeit one may encounter the familiar problem of subject attrition. Of course, this design strategy is not appropriate when a subject's performance in one condition is affected by prior experience with any of the other conditions. As previously mentioned, a within-subject design is inappropriate for studying differential training effectiveness. However, it could be used effectively to investigate the differential suitability of various display formats to enhance target detection where there is little or no differential transfer from display to display. When these within-subject designs are used, caution must be exercised to implement the proper counterbalancing so as to avoid spurious sequence effects.

The within-subject design combines several features of the RSM central-composite design variations previously discussed. Again, a check should be made to insure that the selected value guarantees orthogonality in the case of blocked designs, or rotatability in the case of unblocked designs. The appropriate real-world levels of the experimental factors are then determined accordingly. Data are retained, uncollapsed from repeated observations made at each of the experimental points, thereby affording increased degrees of freedom for the resulting error term. As in the other design variations, the within-subject design permits tests for the significance of blocking and of lack of fit as well as tests of individual partial regression coefficients. In addition, a subject term can be isolated and tested for significance. Because subjects are completely crossed with treatments (every subject receives every treatment once), one can refine the estimate of experimental error variance by accounting for the variability within the individual subjects after assessing the variability within treatment conditions. In a within-subject design the error term which results from merely accounting for the variability of response at the experimental points is comprised of intersubject variations, the interactions between subjects and treatment conditions, and random error. By removing the subject effect a better estimate of experimental error is available for subsequent tests for significance. Moreover, if one assumes no interactions between subjects and treatment conditions, one can test the isolated subject term to determine the existence of significant intersubject variation. (For greater detail concerning the appropriate analysis see Clark and Williges, 1972.)

By way of example, the same four subjects might be observed at each of the 15 experimental points designated in Figure 3, thereby yielding a total of 60 observations. Hypothetical data for such a design are presented in Table4 . Note that the 1.87 value for α is still appropriate because all 15 points, including the centar point, are being replicated an equal number of times as in the between-subjects design with equal replication at all data points. A multiple regression analysis of these hypothetical data yields the following first-order prediction equation:

$$\text{Detection Latency} = 16.44 - 1.16751591\ X_1 - 0.39631381\ X_2 + 0.82118942\ X_3$$

Substituting given levels of display resolution, visual angle, and random noise for X_1, X_2, and X_3, respectively, into this equation provides a corresponding predicted level of detection latency.

- - - - - - - - - - - - - - - - -

Insert Table 4 about here

- - - - - - - - - - - - - - - - -

The results of a subsequent ANOVA performed on the hypothetical data of the regression analysis appear in Table 5. Note the additional "subjects" component into which residual variance has been subdivided. The corresponding degrees of freedom reflect the use of four subjects throughout the experiment. Notice also that the error degrees of freedom are reduced by 3, the degrees of freedom attributed to the subject factor. Had this experiment utilized different subjects throughout, the value of error degrees of freedom would have been 45 rather than 42. But, in the case of within-subject designs, the error term is refined by removing the subject effect from it.

- - - - - - - - - - - - - - - - -

Insert Table 5 about here

- - - - - - - - - - - - - - - - -

Mills and Williges (1972) have used a within-subject design in a recent study of a radar target initiation and maintenance. Their results reveal highly significant intersubject variability which was removed from the regression equation. In addition, the resulting prediction equations appear to demonstrate a high degree of predictive validity to other points within the originally sampled surface (Williges and Mills, 1972).

CONCLUSIONS

The techniques of RSM, and the central-composite design in particular, can be effectively used in human factors research, where the goal is frequently the development of an equation to describe the relationship between human performance and a host of equipment parameters. Certain modifications in the basic

RSM central-composite design, however, appear to make the method more appropriate to research involving human subjects. In making the appropriate design modifications, the investigator must make several major decisions. He must decide whether or not to make repeated observations over a series of experimental points rather than at a single point. If his goal is to develop a global prediction equation to approximate the entire response surface, replication at each of the experimental data-collection points appears to be a wise strategy. The basic central-composite design, calling for replication at only the center point, is perhaps better reserved for preliminary research where the primary aim is to ascertain quickly what major factors appear worthy of more thorough study.

The investigator must also select either a between-subjects or a within-subject design. This choice is dictated by the objectives of his particular experiment. Of the design variants discussed above, those advocating multiple and equal replications at all experimental points, followed by analysis of uncollapsed data, appear the most advantageous, whether they are conceived as between- or within-subject designs. The particular modifications which the investigator elects to implement have ramifications for other aspects of the design such as orthogonal blocking and rotatability. Appropriate adjustments must be made in factor level selection in order to retain such attributes in view of the overall design modification.

ACKNOWLEDGMENTS

This research was supported by the Life Sciences Program, Air Force Office of Scientific Research and is based on a paper given at the Fifteenth Annual Meeting of the Human Factors Society, New York, New York, October 1971. Dr. Glen Finch, Program Manager, Life Sciences Directorate, was the program monitor. The authors wish to express their appreciation to Dr. Richard DeVor and Beverly H. Williges for their helpful comments on an earlier version of this paper.

REFERENCES

Box, G. E. P. and Behken, D. W. Some new three level designs for the study of quantitative variables. Technometrics, 1960, 2, 455-475.

Box, G. E. P. and Hunter, J. S. Multifactor experimental designs for exploring response surfaces. Annals of Mathematical Statistics, 1957, 28, 195-241.

Box, G. E. P. and Wilson, K. B. On the experimental attainment of optimum conditions. Journal of the Royal Statistical Society, Series B (Methodological), 1951, 13, 1-45.

Clark, C. and Williges, R. C. Response surface methodology analyses. Savoy, Ill.: University of Illinois, Institute of Aviation, Aviation Research Laboratory, Technical Report ARL-72-10/AFOSR-72-5, June 1972.

Clark, C., Williges, R. C., and Carmer, S. G. General computer program for response surface methodology analyses. Savoy, Ill.: University of Illinois, Institute of Aviation, Aviation Research Laboratory, Technical Report ARL-71-8/AFOSR-71-1, May 1971.

Cochran, W. G. and Cox, G. M. Experimental designs. (2nd ed.) Chapter 6A. Factorial experiments in fractional replications. New York: Wiley, 1957, 335-375.

Cochran, W. G. and Cox, G. M. Experimental designs. (2nd ed.) Chapter 8A. Some methods for the study of response surfaces. New York: Wiley, 1957, 335-375.

Herzberg, P. A. The parameters of cross-validation. Psychometric Monographs Supplement, 1969, 34, No. 16.

Lord, F. M. and Novick, M. R. Statistical theories of mental test scores. Chapter 13. The selection of predictor variables. Reading, Mass.: Addison-Wesley. 1968, 284-301.

Meyer, D. L. Response surface methodology in education and psychology. The Journal of Experimental Education, 1963, 31, 329-336.

Mills, R. G. and Williges, R. C. Performance prediction in a single-operator simulated surveillance system. Human Factors, 1972, in press.

Simon, C. W. The use of central-composite designs in human factors engineering experiments. Culver City, Calif.: Hughes Aircraft Co., Display Systems and Human Factors Department, Technical Report AFOSR-70-6, December 1970.

Williges, R. C. and Baron, M. L. Transfer assessment using a between-subjects central-composite design. Human Factors, 1972, in press.

Williges, R. C. and Mills, R. G. Predictive validity of central-composite design regression equations. Human Factors, 1972, in press.

Williges, R. C. and North, R. A. Prediction and cross-validation of video cartographic symbol location performance. Human Factors, 1972, in press.

Williges, R. C. and Simon, C. W. Applying response surface methodology to problems of target acquisition. Human Factors, 1971, 13, 511-519.

TABLE 1

Coded Value Coordinates of Data Points for a Second-Order Central-Composite Design in Three Variables

Observation	X_1	X_2	X_3
1	1.0	-1.0	1.0
2	1.0	1.0	-1.0
3	-1.0	1.0	1.0
4	-1.0	-1.0	-1.0
5	-1.0	1.0	-1.0
6	-1.0	-1.0	1.0
7	1.0	-1.0	-1.0
8	1.0	1.0	1.0
9	$-\alpha$	0.0	0.0
10	0.0	$-\alpha$	0.0
11	0.0	0.0	$-\alpha$
12	α	0.0	0.0
13	0.0	α	0.0
14	0.0	0.0	α
15	0.0	0.0	0.0
16	0.0	0.0	0.0
17	0.0	0.0	0.0
18	0.0	0.0	0.0
19	0.0	0.0	0.0
20	0.0	0.0	0.0

TABLE 2

Hypothetical Data in Coded Form for a Three-Factor, Second-Order, RSM Central-Composite Design

		X_1	X_2	X_3	Y
Observation	Block	Resolution	Visual Angle	Random Noise	Detection Latency (Seconds)
1	1	1.00	-1.00	1.00	16.2
2	1	1.00	1.00	-1.00	14.3
3	1	-1.00	1.00	1.00	17.0
4	1	-1.00	-1.00	-1.00	17.4
5	1	0.00	0.00	0.00	15.5
6	1	0.00	0.00	0.00	15.8
7	2	-1.00	1.00	-1.00	16.8
8	2	-1.00	-1.00	1.00	18.1
9	2	1.00	-1.00	-1.00	14.9
10	2	1.00	1.00	1.00	16.2
11	2	0.00	0.00	0.00	15.0
12	2	0.00	0.00	0.00	14.8
13	3	-1.63	0.00	0.00	19.0
14	3	0.00	-1.63	0.00	17.3
15	3	0.00	0.00	-1.63	14.8
16	3	1.63	0.00	0.00	13.9
17	3	0.00	1.63	0.00	14.6
18	3	0.00	0.00	1.63	19.2
19	3	0.00	0.00	0.00	15.8
20	3	0.00	0.00	0.00	15.7

TABLE 3

First-Order Regression Analysis of Variance Summary Table for Hypothetical Detection Latency Data

Source	df	MS	F
Regression	(3)	10.73	536.50**
b_1	1	19.26	963.00**
b_2	1	3.37	168.51**
b_3	1	9.54	477.00**
Residual	(16)	0.71	
Blocks	2	0.21	10.50*
Lack of Fit	11	0.99	49.50**
Error	3	0.02	
Total	(19)		

* $p < .05$

** $p < .001$

Multiple Regression Coefficient, R, = 0.86

Coefficient of Determination, R^2, = 0.74

TABLE 4

Hypothetical Data in Coded Form for a Three-Factor, Second-Order, RSM Central-Composite Design Using Repeated Measures on Four Subjects

			Detection Latency (Seconds) For Four Subjects			
Resolution	Visual Angle	Random Noise	S_1	S_2	S_3	S_4
1.00	-1.00	1.00	15.8	15.9	16.1	16.4
1.00	1.00	-1.00	14.3	14.5	14.0	14.8
-1.00	1.00	1.00	17.0	17.3	17.1	16.9
-1.00	-1.00	-1.00	17.4	17.5	17.0	17.3
-1.00	1.00	-1.00	16.8	16.7	17.0	17.0
-1.00	-1.00	1.00	18.1	18.3	18.6	18.1
1.00	-1.00	-1.00	14.9	15.2	14.5	15.0
1.00	1.00	1.00	16.2	16.7	16.4	15.9
-1.87	0.00	0.00	19.0	19.1	18.9	19.5
0.00	-1.87	0.00	17.3	16.9	17.4	16.8
0.00	0.00	-1.87	15.1	15.3	14.4	15.0
1.87	0.00	0.00	13.9	14.2	13.7	14.1
0.00	1.87	0.00	14.9	15.0	14.8	15.0
0.00	0.00	1.87	19.2	19.0	20.0	18.9
0.00	0.00	0.00	15.8	16.1	16.4	16.0

TABLE 5

First-Order Regression Analysis of Variance Summary Table for Hypothetical Detection Latency Data of Four Subjects

Source	df	MS	F
Regression	(3)	43.87	548.37**
b_1	1	81.75	1021.87**
b_2	1	9.42	117.75**
b_3	1	40.44	505.50**
Residual	(56)	0.42	
Blocks	2	0.65	8.13*
Subjects	3	0.05	0.63
Lack of Fit	9	2.04	25.50**
Error	(42)	0.08	
Total	59		

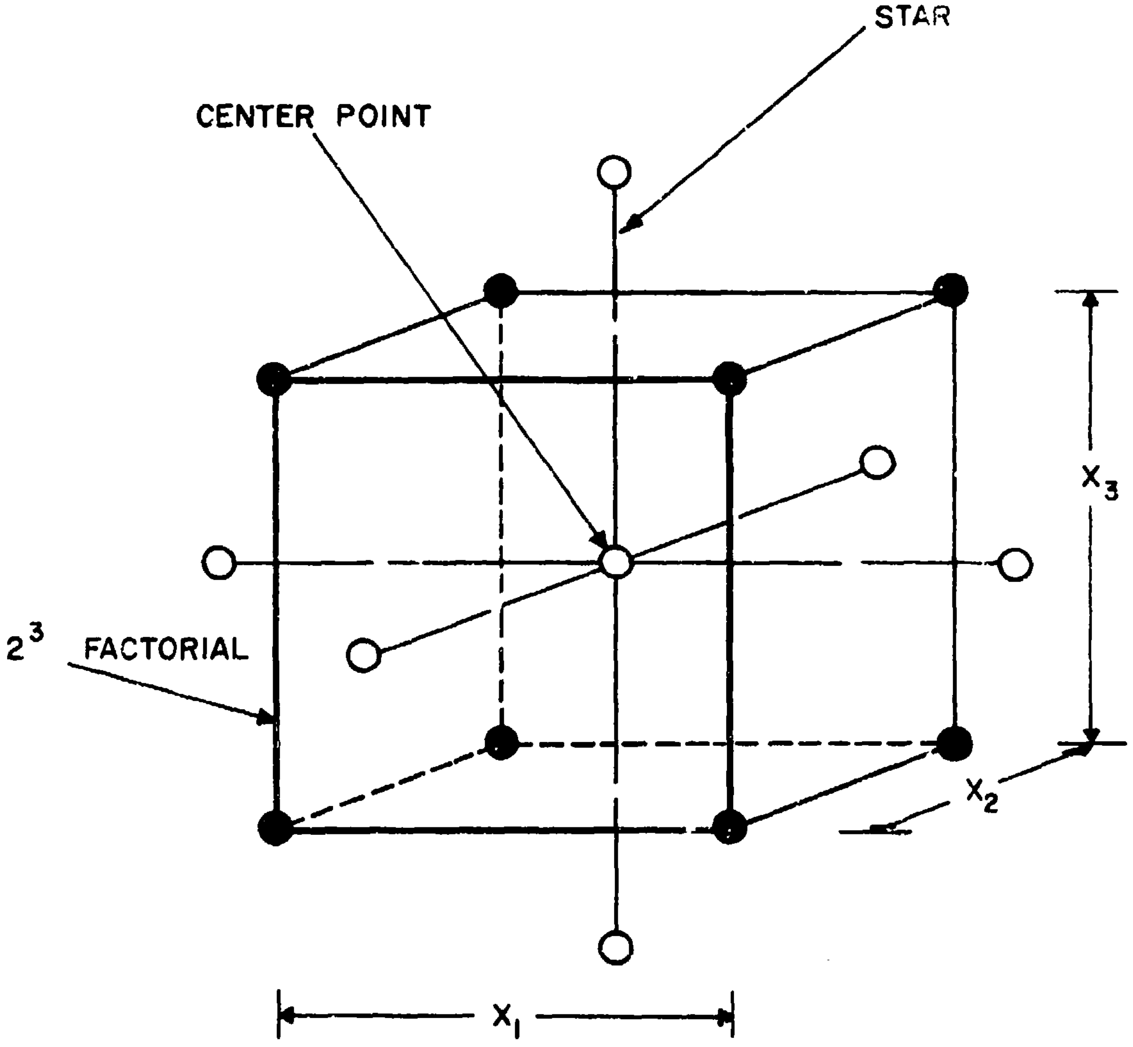
* $p < .01$

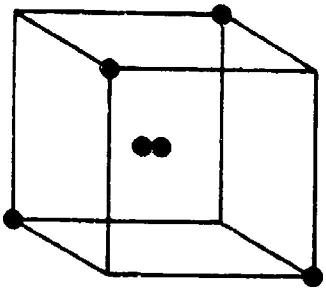
** $p < .00i$

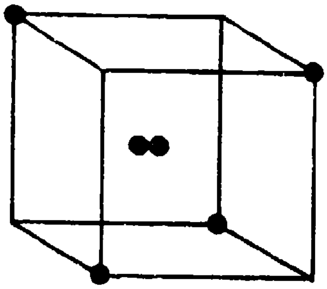
Multiple Regression Coefficient, R, = 0.92

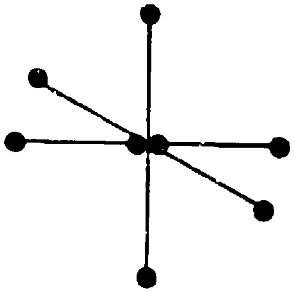
Coefficient of Determination, R^2, = 0.85

LIST OF FIGURES

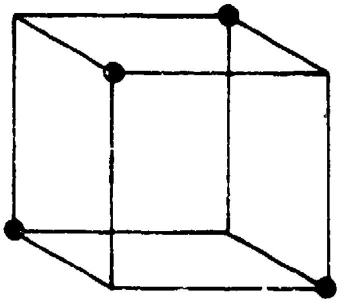

STAR
CENTER POINT
2^3 FACTORIAL
X_3
X_2
X_1

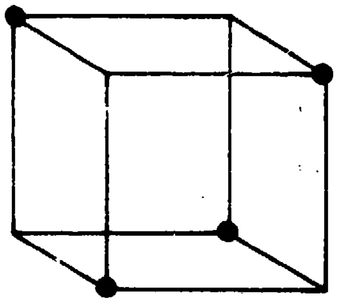


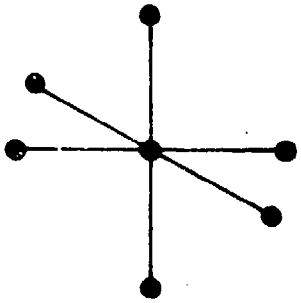




BLOCK 1	BLOCK 2	BLOCK 3
(1, -1, 1)	(-1, 1, -1)	(-1.633, 0, 0)
(1, 1, -1)	(-1, -1, 1)	(0, -1.633, 0)
(-1, 1, 1)	(1, -1, -1)	(0, 0, -1.633)
(-1, -1, -1)	(1, 1, 1)	(1.633, 0, 0)
(0, 0, 0)	(0, 0, 0)	(0, 1.633, 0)
(0, 0, 0)	(0, 0, 0)	(0, 0, 1.633)
		(0, 0, 0)
		(0, 0, 0)







BLOCK 1	BLOCK 2	BLOCK 3
(1, -1, 1)	(-1, 1, 1)	(-1.87, 0, 0)
(1, 1, -1)	(-1, -1, 1)	(0, -1.87, 0)
(-1, 1, 1)	(1, -1, -1)	(0, 0, -1.87)
(-1, -1, -1)	(1, 1, 1)	(1.87, 0, 0)
		(0, 1.87, 0)
		(0, 0, 1.87)
		(0, 0, 0)

BIOGRAPHY

CHRISTINE CLARK (Response Surface Methodology Central-Composite Design Modifications for Human Performance Research) is a graduate student in psychology at the University of Illinois majoring in social psychology and minoring in measurement. She received her B.A. in mathematics from Illinois in 1969 and entered the Ph.D. Program in the Psychology Department that fall. As a first year graduate student she held a traineeship in measurement, granted by the United States Public Health Service. During the last two years she has had a research assistantship at the Aviation Research Laboratory of the Institute of Aviation. Her primary concerns have been the investigation of the utility of response surface methodology as a research tool and the coincident development of computer procedures to analyze related data.

BIOGRAPHY

ROBERT C. WILLIGES (Response Surface Methodology Central-Composite Design Modifications for Human Performance Research) is the Assistant Head for Research of the Aviation Research Laboratory of the Institute of Aviation and Associate Professor of Psychology and of Aviation at the University of Illinois. He received his A.B. degree in psychology from Wittenberg University in 1964 and his M.A. and Ph.D. degrees in engineering psychology from The Ohio State University in 1966 and 1968, respectively. While at Ohio State he was a research assistant at the Human Performance Center and conducted research on team training and monitoring of complex computer-generated displays. Prior to joining the Aviation Research Laboratory in 1970, he was Assistant Director of the Highway Traffic Safety Center at the University of Illinois. His current research interests include problems of visual monitoring performance, inspector behavior, and human performance in complex system operation including investigation of rate-field, frequency-separated, and visual time-compressed displays, interpretability of TV-displayed cartographic information, transfer of training, and applications of response surface methodology.

Transfer Assessment Using a Between-Subjects Central-Composite Design

ROBERT C. WILLIGES and MARVIN L. BARON[1], University of Illinois at Urbana-Champaign

Transfer of training from a pursuit rotor to an epicycloid pursuit rotor was assessed by means of a Response Surface Methodology (RSM) central-composite design. Number of training trials, time between training trials, and tracking speed of the training task were combined in a three-factor, RSM central-composite design. Multiple regression prediction equations relating these three independent variables to trials to criterion on the epicycloid pursuit rotor were calculated for both an unreplicated and replicated RSM design. A representative first-order response surface was plotted for the replicated design. The results were discussed in terms of necessary RSM central-composite design modifications and the overall applicability of using RSM in transfer of training research.

INTRODUCTION

With the development of Response Surface Methodology (RSM) by Box and Wilson (1951), an experimental technique was introduced that specifies procedures for the economical collection of data in multiparameter research. Although RSM was originally developed as a series of experimental steps to ascertain the optimum combination of variubles for producing maximum yield of a chemical process, the experimental design procedures are applicable to human performance research. One aspect of RSM that appears to be particularly useful is the central-composite design. This design is often used in the systematic exploration of complex response surfaces. Because of the economy and efficiency of the central-composite design (see Williges and Simon, 1971), it may be useful in determining an overall multiple regression prediction equation which describes the combined relationship among several independent variables in producing a certain level of performance.

Clark and Williges (1972a) suggested various modifications that make RSM central-composite designs more applicable to human performance research. The major purposes of this study are to investigate one of the proposed design modifications concerning data replication and to use a between-subjects RSM central-composite design in predicting the simultaneous effects of several variables affecting transfer of training by means of a single multiple regression equation.

Although RSM has been used in engineering for many years, only one limited application has been made to problems of human learning. Meyer (1963) used RSM to study the effects of four factors on the amount of retroactive inhibition induced in a typical retroactive inhibition paradigm in verbal learning. A response surface was plotted relating amount of recall to variation in the independent variables, and the point of maximum recall was determined.

One major goai of any training program is to maximize positive transfer. Many task dimensions, such as distribution of practice, degree of original learning, and task difficulty, have been investigated to determine their significance in producing transfer. The separate effects of these variables are well documented in the research literature, but little research has been concerned with the combined effects of these variables. In any training situation, however, all of these variables are operating together, and their particular combination determines the actual amount of positive transfer. To understand the underlying relationships of these variables, it is important to investigate all of the significant variables simultaneously.

Distribution of practice is a dimension that has been extensively investigated in the context of transfer of training. Digman (1959) demonstrated that performance under massed practice may appear to be depressed when compared to distributed practice, although it does not affect learning of a motor skill. Studies by Reynolds and Adams (1953) and Denney, Frisbey, and Weaver (1955) have shown that if subjects are trained under massed practice and then transferred to distributed practice, their performance improves to the level of control subjects tracking solely with distributed practice. Massed practice, therefore, tends to depress the standard of performance rather than the rate of learning.

The results of studies dealing with the degree of original learning on transfer are straightforward: positive transfer increases as a function of the amount of original learning. To summarize the effect of practice, Mandler (1962) states that a small amount of practice produces an initial negative transfer, then transfer returns to zero with more practice, and finally positive transfer occurs with additional practice. Studies by Siipola and Israel (1933) and Mandler and Heinemann (1956) provide support for this contention. Simply stated, negative transfer has the greatest likelihood of occurring after relatively little practice on the original task.

Unfortunately, the relationship between task difficulty and transfer is not as simple. In some cases, transfer is greater from a difficult to an easy task, and sometimes the reverse is true. Generalizations about the effect of task difficulty upon transfer are limited because so many different tasks have been used to study the effect of this dimension, and it is not easy to determine what constitutes comparable levels of difficulty with different tasks (Day, 1956).

An attempt to explain the findings of differential transfer resulting from variations in task difficulty has been made by Holding (1965). His principle of "inclusion" states that if the requirements of a subsequent transfer task are contained in the training task, transfer performance will be high. When inclusion of these requirements is not present, transfer will be low. When the inclusion principle applies to a task, one would expect to find greater transfer from the difficult-to-easy direction because the difficult training task contains the skill components required for mastery of the easy transfer task.

Holding also offered an explanation for differential transfer favoring the easy-to-difficult order of tasks by proposing his hypothesis of "performance standards." He states that a subject develops high performance standards when working with an easy task. Good performance on the transfer task will result when these high standards are carried over to the more difficult transfer task.

By using an experimental task similar to that used in previous research, earlier experimental results of variables representing dimensions of amount of original learning, task difficulty, and distribution of practice can be used as a comparative baseline. The results of the subsequent RSM central-composite prediction equation can be readily compared to this baseline to ascertain compatibility of results.

METHOD

Apparatus

A pursuit rotor (Melton, 1947) was used as the training task, and an epicycloid pursuit rotor (Barch and Lewis, 1951) was used as the transfer task. A small brass target, 1/2 inch in diameter, moved clockwise on a rotating disc in the pursuit rotor circumscribing a 12-inch circular path. Although the identical target size was used in the epicycloid pursuit, the target path was heart-shaped rather than circular. This path was generated by a small satellite disc that revolved about a point 3 1/2 inches from the center of the large disc. During each clockwise rotation of the large disc, the satellite revolved once in the same direction.

A spring-loaded metal stylus was used to track the target on both the pursuit rotor and the epicycloid pursuit rotor. Time-on-target was recorded to the nearest second by means of a clock-timer.

Experimental Design

A three-factor, second-order RSM central-composite design was used. According to the design, five levels of each factor were needed with the coded values, -1.633, -1, 0, +1, +1.633. A 2^3 factorial design was constructed from the +1 and -1 coded values, and a 2·3 star component was constructed from the values, +1.633 and -1.633. The design was blocked across three different experimenters to control against any experimenter bias. A coded α value of 1.633 was chosen to maintain orthogonal blocking. The various coded data points collected by each experimenter during a single replication of the RSM design are listed in Table 1. The complete replication of the RSM central-composite design included 20 data points, 6 of which were collected by Experimenter 1, 6 by Experimenter 2, and 8 by Experimenter 3. Table 1 also shows that the center point (0, 0, 0) was observed twice in each block in order to obtain an estimate of experimental error. Note that the design was blocked such that Experimenters 1 and 2 each collected data on a one-half replicate of the 2^3 factorial design, and Experimenter 3 collected data on the star component of the design. The third-order interaction was chosen as the defining relationship for the one-half replicates so that no first- or second-order components would be confounded with experimenters or each other in the second-order RSM central-composite design. (See Box and Wilson, 1951; Simon, 1970; and Clark and Williges, 1972b for additional details concerning the central-composite design.)

- - - - - - - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - - - - - - -

The three factors were amount of original learning, task difficulty, and distribution of practice during training. Amount of original learning was manipulated in terms of the number of training trials with actual values of 5, 11, 20, 29, and 35 trials for the coded values of -1.633, -1, 0, +1, and +1.633, respectively. Task difficulty was represented by the tracking speed of the pursuit rotor during training with actual values of 5, 26, 60, 94, and 115 r.p.m. Distribution of practice was varied by changing the time between training trials with actual values of 15, 27, 45, 63, and 75 seconds.

Subjects

A total of 40 subjects were selected from students enrolled in the primary flight training course at the University of Illinois and from students currently holding an FAA private pilot certificate. Flight students and private pilots were used to obtain a group of subjects with more homogeneous perceptual-motor abilities than subjects from the general population. Twenty subjects were used in each of two replications of the design. Each subject was paired with another subject receiving the same experimental training condition. The subject in each pair requiring the fewer trials to reach criterion during transfer was awarded one hour of airplane rental time. The other subject in each pair received no reward for his participation.

Procedure

Each subject received the appropriate combination of the three independent variables during training on the pursuit rotor. Trials were 60 seconds in length. The next day, each subject transferred to the epicycloid pursuit rotor. Before beginning the transfer task, each was shown a diagram of the heart-shaped path of the target. Each subject was required to continue tracking the epicycloid pursuit rotor until he attained a criterion of at least 10 seconds on target during two successive 60-second trials. The transfer task consisted of the center levels of both tracking speed and time between trials used during training, namely, 45 r.p.m. and 60 seconds.

RESULTS

The results were analyzed in two different stages. First, the data were analyzed as a traditional, RSM central-composite design with multiple observations at only the center point of the design. Second, the complete design was replicated, and the data were analyzed by considering multiple observations at each point in the design. A computer program developed by Clark, Williges, and Carmer (1971) was used to conduct the RSM regression analyses during both stages. A detailed discussion of these specific calculation procedures is presented by Clark and Williges (1972a).

Unreplicated Design

Using the data obtained from the 20 treatment conditions, a complete first-order standard multiple regression equation was obtained using the following correlation matrix solution:

$$\underline{b}' = [r_{X_iX_k}]^{-1} [r_{X_iY}] , \qquad (1)$$

where $\underline{b}'$ is a column vector of the m standard partial regression coefficients b_i', $i = 1, m$; $[r_{X_iX_k}]^{-1}$ is the inverse of the m x m correlation matrix, the elements of which are all pairwise correlations between the m independent variables; and $[r_{X_iY}]$ is the column vector, the elements of which are the pairwise correlations between Y and each of the m independent variables. In the case of a complete three-factor first-order equation, m is three.

The three resulting standard partial regression coefficients, b_i', $i = 1, 3$, of Equation 1 are readily converted to the corresponding nonstandard coefficients, b_i, according to the following relation:

$$b_i = b_i' \frac{s_Y}{s_{X_i}} . \qquad (2)$$

The intercept value, b_0, is obtained as follows:

$$b_0 = \overline{Y} - b_1\overline{X}_1 , \ldots , - b_i\overline{X}_i . \qquad (3)$$

The resulting nonstandard, complete first-order multiple regression for these data would be in the form

$$Y = b_0 + b_1X_1 + b_2X_2 + B_3X_3 . \qquad (4)$$

Specifically, the resulting multiple regression equation using the uncoded data was:

$$\text{Trials to Criterion} = 47.18 - 0.38\,N - 0.08\,T - 0.39\,S$$

where Trials to Criterion = two successive 60-second transfer trials on the epicycloid pursuit with at least 10 seconds on target on each trial; N = the number of training trials on the pursuit rotor; T = time between training trials; and S = the tracking speed of the pursuit rotor. The multiple correlation coefficient was .68.

The regression analysis can subsequently be submitted to an analysis of variance to estimate the reliability of the various effects. Essentially, the total variation is partitioned into regression sum of squares (SS) and residual SS. Regression can be further subdivided into the additional SS due to each partial regression coefficient. Likewise, residual SS in this analysis can be partitioned into replication SS (error), lack of fit SS, and experimenter SS. The general equations for calculating these effects are as follows:

$$\text{Total SS} = \Sigma Y_i^2 + (\Sigma Y_i)^2/N \; ; \qquad (5)$$

$$\text{Regression SS} = \underline{b}^t \underline{g} \; , \qquad (6)$$

where $\underline{b}^t$ is the row vector transpose of the column vector of partial regression coefficients, and $\underline{g}$ is the column vector of corrected cross products between the dependent variable and the various independent variables;

$$\text{Residual SS} = \text{Total SS} - \text{Regresion SS} \; ; \qquad (7)$$

$$\text{additional SS due to } X_i = b_i^2/c_{ii} \; , \qquad (8)$$

where b_i is the i^{th} partial regression coefficient and c_{ii} is the element occupying the i^{th} row and i^{th} column of the inverse of the corrected sum of squares cross-product matrix;

$$\text{Experimenter SS} = \sum_{k=1}^{NE} m_{E_k} (\overline{Y} - \overline{Y}_{E_k})^2 \; , \qquad (9)$$

where $\overline{Y}$ is the grand mean of the dependent variables across all observations, $\overline{Y}_{E_k}$ is the mean of the dependent variables across the observations comprising the k^{th} experimenter, m_{E_k} is the number of observations comprising the k^{th} experimenter, and NE is the number of experimenters comprising the entire design;

$$\text{Replication SS} = \sum_i (Y_i - \overline{Y}_r)^2 \quad , \qquad (10)$$

where the index i corresponds to the repeated observations at the center point (0, 0, 0) and $\overline{Y}_r$ is the mean of the dependent variable across the replications of the center point. This value is calculated separately for replications under each experimenter and then summed across experimenters;

and

$$\text{Lack of Fit SS} = \text{Residual SS} - \text{Experimenter SS} - \text{Replications SS}. \qquad (11)$$

The center portion of Table 2 summarizes the results of a subsequent analysis of variance performed on the regression analysis. Using replications at the six center points of the RSM design as an estimate of error, the analysis yielded nonsignificant effects due to regression, partial regression weights, experimenters, and lack of fit ($\underline{p} > .10$).

- - - - - - - - - - - - - - - - -

Insert Table 2 about here

- - - - - - - - - - - - - - - - -

Because error was estimated only in the center of the design yielding three degrees of freedom, the error variance was large and resulted in the other effects not being statistically reliable. If the entire design were replicated, a more sensitive estimate of error could be obtained because of the substantial increase in the degrees of freedom of the error. This procedure would seem to be particularly necessary in a between-subjects design assessing human performance on a perceptual-motor task where large individual differences might be expected. Consequently, the entire RSM central-composite design was replicated, thereby adding an additional 20 observations to the experiment.

<u>Replicated Design</u>

The first-order, RSM multiple regression prediction equation for uncoded, replicated data was:

$$\text{Trials to Criterion} = 47.74 - 0.36\,N - 0.06\,T - 0.40\,S$$

The right portion of Table 2 shows the analysis of variance for the replicated multiple regression equation. The regression equation now accounts for a significant amount of the variability even though the multiple correlation coefficient remains approximately the same as the unreplicated data (R = .69). In addition, the analysis of variance demonstrates that number of training trials and tracking speed of the training task were both significant contributors to prediction of trials to criterion during transfer. Time between training trials, however, was not a significant predictor ($\underline{p} > .10$), and the lack of fit was not significant ($\underline{p} > .05$). Note that the degrees of freedom contributed by the additional 20 points in the replicated design all appear in the replication term, thereby providing a more sensitive estimate of error.

Figure 1 depicts the linear response surface defined by the replicated design regression equation. The two plotted curves on the graph indicate transfer performance in terms of 15 and 25 transfer trials to reach criterion. The transfer surface is primarily a function of the number of training trials and tracking speed of the training task. Time between training trials affects the contour of the transfer response surface only slightly. In addition to plotting the transfer surface, these curves also illustrate the tradeoffs that must be made among the independent variables in order to obtain a given number of trials to criterion on the transfer task.

Insert Figure 1 about here

DISCUSSION

After comparing the results of the replicated and the unreplicated design, it is clear that RSM designs need to be modified somewhat when applied to human performance. Although the resulting prediction equations were similar in both the replicated and unreplicated designs, the replicated design was more sensitive. When different subjects are used in a motor skills task, the results of this study indicate that the between-subject variability is such that replication is desirable over the entire design.

It should be noted that it is not necessary to replicate the entire design. The design can be replicated with only 2 center points rather than with the 12 required for a complete replication of the intact three-factor, RSM central-composite design. When blocking is used, an adjustment must be made in the coded value of the noncenter points (α) of the third block in order to maintain orthogonality between the block effects and the independent variables. Procedures for calculating this adjusted value are provided by Cochran and Cox (1957) and Clark and Williges (1972b).

The effects of all three independent variables used in the replicated multiple regression prediction equation appear to be compatible with previous research. As the number of training trials or the degree of original learning increases, trials to criterion in transfer decrease. Ellis (1965) states that positive transfer increases with increasing practice on the training task.

Time between trials was an unreliable predictor in this study; but the trend suggests that the longer the time between trials, the better the performance on the transfer task. This result is consistent with findings resulting in better performance with distributed rather than massed practice (Digman, 1959). It is not altogether surprising that time between trials was not a significant contributor to transfer in view of previous research in perceptual-motor skill that suggests this variable primarily affects performance rather than learning (Reynolds and Adams, 1953).

Tracking speed was a strong determiner of transfer. Because trials to criterion decreased as the tracking speed of the training task increased, the effect of this variable is in line with the point of view which contends that higher transfer results from the shift from a difficult to an easy task. This result appears to support the "inclusion" principle of Holding (1965), because the transfer task consisted of a track involving a continuously changing rate of rotation. To the extent that the training task included the higher rates of tracking during training, transfer performance was improved.

Although these results support previous research, the real value of this study is that it provides a simultaneous investigation of all three variables, thereby providing information as to the relative importance of each. Obviously, the relationship among these variables cannot be extended beyond the limits of the range of levels tested. Tracking speed during training could be increased to a point where the subjects could no longer track the target. Similarly, although transfer is a positive function of the number of training trials, a point will be reached beyond which additional trials will no longer produce a significant increase in transfer. Consequently, one would expect the transfer surface to become nonlinear as the range of variables increases.

Even in the present results, there is some indication of nonlinear or higher-order effects. The lack of fit in the replicated design in Table 2 was not significant at the .05 level. If the alpha error is increased to .10 to reduce the probability of a beta error, the lack of fit becomes significant. A subsequent multiple regression analysis fitting complete first-order (linear) and second-order (quadratic) tenns with the coded data yielded no significant second-order effects. The lack of fit of the complete second-order analysis was significant ($\underline{p} < .05$), however, suggesting that still higher-order terms may be present.

The results of this study clearly indicate that RSM techniques provide both a useful and economical approach for investigating the effects of several variables on human transfer performance. Although this initial study demonstrates the potential of the technique and includes representative equipment and procedural variables of recognized importance in transfer, additional research that includes other variables and more complex perceptual-motor tasks is necessary.

ACKNOWLEDGMENTS

Initial work on this study was supported by a grant from the Link Foundation. Additional support was provided by the Life Sciences Program, Air Force Office of Scientific Research. Dr. Glen Finch, Program Manager, Life Sciences Directorate, was the program monitor for AFOSR. This article is based on a paper given at the Fifteenth Annual Meeting of the Human Factors Society, New York, New York, October 1971. The authors wish to express their appreciation to Fuat Ince and Robert A. North who assisted in the data collection phase of the experiment and to Beverly H. Williges for her helpful suggestions on earlier drafts of this paper.

REFERENCES

Barch, A. N. and Lewis, D. A demonstration of retroactive interference in pursuit rotor learning. Port Washington, N. Y.: Special Devices Center, Technical Report No. 166-00-1, December 1951.

Box, G. E. P. and Wilson, K. B. On the experimental attainment of optimal conditions. Journal of the Royal Statistical Society, Series B (Methodological), 1951, 13, 1-45.

Clark, C. and Williges, R. C. Response surface methodology analyses. Savoy, Ill.: University of Illinois, Institute of Aviation, Aviation Research Laboratory, Technical Report ARL-72-10/AFOSR-72-5, June 1972. (a)

Clark, C. and Williges, R. C. Response surface methodology central-composite design modifications for human performance research. Human Factors, 1972, in press. (b)

Clark, C., Williges, R. C., and Carmer, S. General computer program for response surface methodology analyses. Savoy, Ill.: University of Illinois, Institute of Aviation, Aviation Research Laboratory, Technical Report ARL-71-8/AFOSR-71-1, May 1971.

Cochran, W. G. and Cox, G. M. Some methods for the study of response surfaces. Experimental designs. New York: Wiley, 1957, 335-375.

Day, R. H. Relative task difficulty and transfer of training in skilled performance. Psychological Bulletin, 1956, 53, 160-168.

Denney, M. R., Frisbey, N., and Weaver, J., Jr. Rotary pursuit performance under alternate conditions of distributed and massed practice. Journal of Experimental Psychology, 1955, 49, 48-54.

Digman, J. M. The growth of a motor skill as a function of practice. Journal of Experimental Psychology, 1959, 57, 310-316.

Ellis, H. The transfer of learning. New York: Macmillan, 1965.

Holding, D. H. Principles of training. London: Pergamon Press, 1965.

Mandler, G. From association to structure. Psychological Review, 1962, 69, 415-427.

Mandler, G. and Heinemann, S. H. Effect of overlearning of a verbal response on transfer of training. *Journal of Experimental Psychology*, 1956, **51**, 39-46.

Melton, A. W. *Apparatus tests*. U. S. Army Air Forces Aviation Psychology Program, Report No. 4, 1947.

Meyer, D. L. Response surface methodology in education and psychology. *Journal of Experimental Education*, Summer 1963, **31** (4), 330-336.

Reynolds, B. and Adams, J. A. Effect of distribution and shift in distribution of practice within a single training session. *Journal of Experimental Psychology*, 1953, **46**, 137-145.

Simon, C. W. Use of central composite designs in human factors engineering experiments. Culver City, Calif.: Hughes Aircraft Co., Display and Human Factors Department, Technical Report AFOSR-70-6, December 1970.

Siipola, E. M. and Israel, H. E. Habit interference as dependent upon stage of training. *American Journal of Psychology*, 1933, **45**, 205-227.

Williges, R. C. and Simon, R. C. Applying response surface methodology to problems of target acquisition. *Human Factors*, 1971, **13**, 511-519.

FOOTNOTE

[1] Now at the U.S. Army Electronics Command, Avionics Laboratory, Environmental Sensing and Instrumentation Technical Area, Fort Monmouth, New Jersey.

TABLE 1

Coded Data Points of the RSM Central-Composite Design

Treatment Condition	Experimenter	Training Trials	Time Between Trials	Tracking Speed
1	1	-1	-1	1
2	1	1	-1	-1
3	1	-1	1	-1
4	1	1	1	1
5	1	0	0	0
6	1	0	0	0
7	2	-1	-1	-1
8	2	1	-1	1
9	2	-1	1	1
10	2	1	1	-1
11	2	0	0	0
12	2	0	0	0
13	3	-1.633	0	0
14	3	1.633	0	0
15	3	0	-1.633	0
16	3	0	1.633	0
17	3	0	0	-1.633
18	3	0	0	1.633
19	3	0	0	0
20	3	0	0	0

TABLE 2

First-Order Regression Analysis of Variance Summary Table of Unreplicated and Replicated RSM Central-Composite Designs

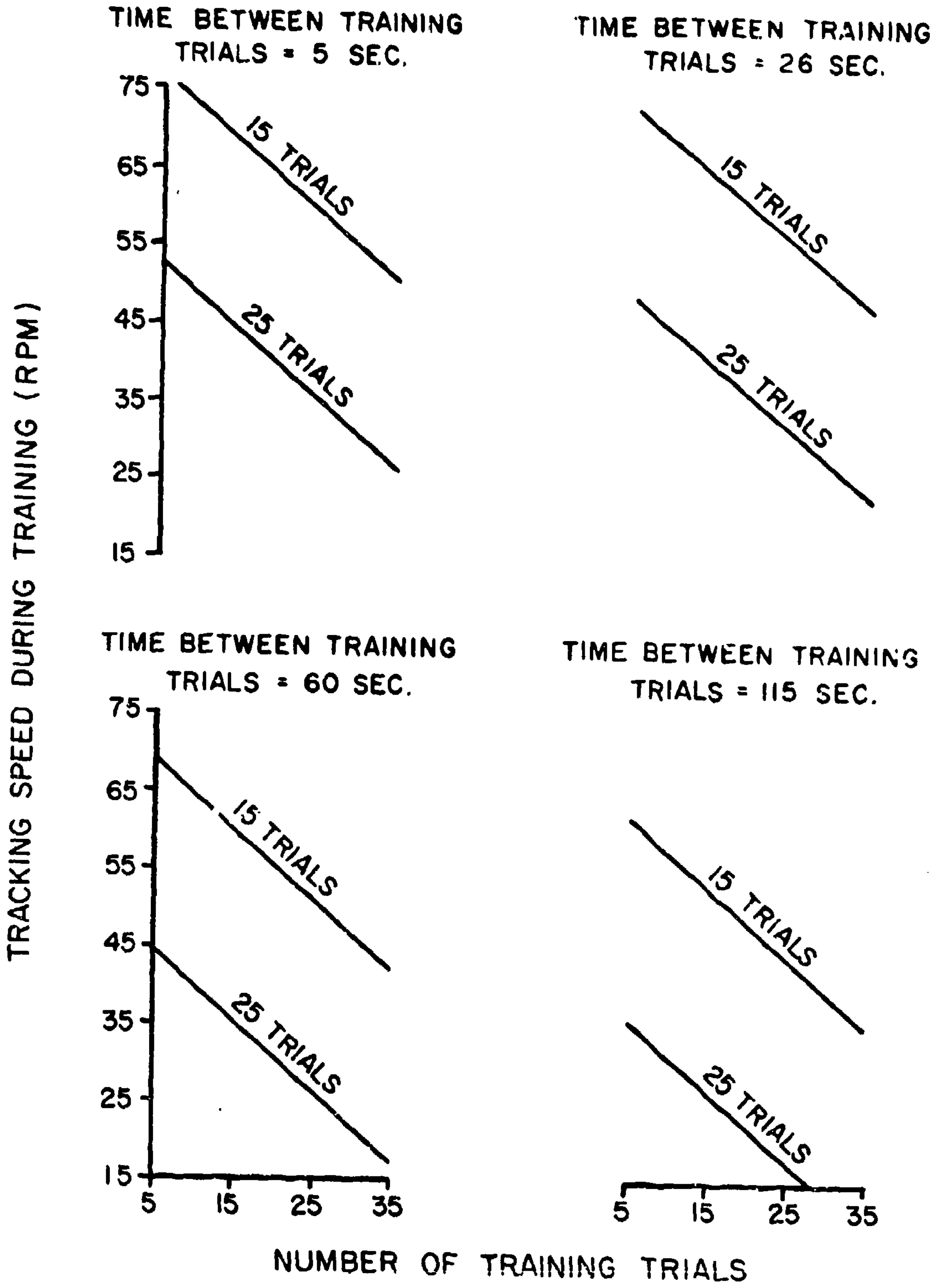
	Unreplicated Design			Replicated Design		
Source	df	MS	F	df	MS	F
Regression	(3)	301.32	2.21	(3)	592.57	14.08**
Number of Training Trials	1	156.10	1.15	1	281.33	6.69*
Time Between Trials	1	96.02	-	1	98.61	2.34
Tracking Speed	1	651.84	4.78	1	1397.77	33.22**
Residual	(16)	64.81		(36)	53.70	
Experimenters	2	41.31	-	2	11.35	-
Lack of Fit	11	49.58	-	11	85.71	2.04
Replications[a]	3	136.33		23	42.08	
Total	(19)			(39)		

[a] Error term

* $p < .05$

** $p < .01$

LIST OF FIGURES


TIME BETWEEN TRAINING TRIALS = 5 SEC.
TIME BETWEEN TRAINING TRIALS = 26 SEC.
TIME BETWEEN TRAINING TRIALS = 60 SEC.
TIME BETWEEN TRAINING TRIALS = 115 SEC.
TRACKING SPEED DURING TRAINING (RPM)
75
65
55
45
35
25
15
15 TRIALS
25 TRIALS
5
15
25
35
NUMBER OF TRAINING TRIALS

BIOGRAPHY

ROBERT C. WILLIGES (Transfer Assessment Using a Between-Subjects Central-Composite Design) is the Assistant Head for Research of the Aviation Research Laboratory of the Institute of Aviation and Associate Professor of Psychology and of Aviation at the University of Illinois. He received his A.B. degree in psychology from Wittenberg University in 1964 and his M.A. and Ph.D. degrees in engineering psychology from The Ohio State University in 1966 and 1968, respectively. While at Ohio State he was a research assistant at the Human Performance Center and conducted research on team training and monitoring of complex computer-generated displays. Prior to joining the Aviation Research Laboratory in 1970, he was Assistant Director of the Highway Traffic Safety Center at the University of Illinois. His current research interests include problems of visual monitoring performance, inspector behavior, and human performance in complex system operation including investigation of rate-field, frequency-separated, and visual time-compressed displays, interpretability of TV-displayed cartographic information, transfer of training, and applications of response surface methodology.

BIOGRAPHY

MARVIN L. BARON (Transfer Assessment Using a Between-Subjects Central-Composite Design) received a B.S. in psychology from Brooklyn College in 1963, and an M.A. in engineering psychology from the University of Illinois in 1966. While at Illinois, he was a Research Assistant at the Aviation Psychology Laboratory and conducted research in verbal learning and mediation. His M.A. thesis dealt with the relationship between motion acuity and static acuity in a simulated driving situation. After receiving his M.A. in 1966, he taught for the New York City Board of Education. In 1967, he accepted a commission in the Medical Service Corps as a research psychologist and was assigned to the Skills Analysis Branch of the U. S. Army Medical Research Laboratory, Fort Knox, Kentucky where he performed research in the analysis of complex motor skills and skill decrement, development of motor skills tasks, transfer of training, and the analysis of skills involved in vehicle driving. He returned to the University of Illinois in the fall of 1970 to continue his doctoral studies and conducted research dealing with transfer of training assessment by means of Response Surface Methodology and the transfer effectiveness of driving simulation. In the spring of 1972, he accepted a position at the U.S. Army Avionics Laboratory where he is presently concerned with the evaluation of cockpit instrumentation and display-control systems for use in rotary wing aircraft.

Prediction and Cross-Validation of Video Cartographic Symbol Location Performance

ROBERT C. WILLIGES and ROBERT A. NORTH, University of Illinois at Urbana-Champaign

A Response Surface Methodology central-composite design was used to obtain multiple regression prediction equations of performance on a video cartographic symbol search task. Observers were required to locate the position of designated target symbols on a series of maps displayed on black and white and color television (TV) monitors. The variables used to predict both location and latency performance were focus, density of nontarget symbols, visual angle of the observer, and TV raster lines per mm of actual map area. Prediction equations were compared for black and white and color TV monitors through collapsed and uncollapsed, within-subject data analyses. Both analysis procedures were compared in terms of resulting sensitivity and in terms of the predictive validity of the regression equations as determined in cross-validation. It was concluded that the uncollapsed, within-subject designs provided the better prediction equations.

INTRODUCTION

The rapid world-wide dissemination of current cartographic information may be facilitated by transmitting newly updated cartographic images by television (TV). For TV displays to be used effectively for this purpose, the systems designer must know the relationships between various display and situational variables and image interpretability. By knowing the simultaneous effects of these variables, presented in the form of performance prediction equations, the designer can make meaningful tradeoffs among the many variables operating in the system.

One method of predicting performance is to develop a theoretical model describing the simultaneous effects of various variables of interest. An attempt to incorporate several parameters into a predictive model of observer performance was undertaken by Greening and Wyman (1970). The model is based upon a series of probabilities associated with several variables in the task and represents the culmination of several years of research on each of these variables. Although the predictive validity of the model is reportedly high, the factors of time and cost in developing such a model are the difficulties with this approach. In addition, certain assumptions must be made to evaluate the various parameters used in the model.

An alternative approach to theoretical model building would be to derive an empirical multiple regression equation which predicts observer performance as a weighted combination of the specific display and situational variables of interest. Regression equations are easily obtained,and the experimenter need only collect enough data to solve for the various parameters of his regression model. For his resulting prediction equation to have high predictive validity, however, the experimenter must derive his prediction equation from a sample of data that adequately represents the range and relationships of the variables of interest.

Williges and Simon (1971) pointed out that certain Response Surface Methodology (RSM) procedures as originally developed by Box and Wilson (1951) may provide economical and efficient techniques of collecting data for deriving multiple regression prediction equations. In particular, the central-composite design appears quite useful for this purpose.

This paper illustrates the use of a within-subject, RSM central-composite design to develop multiple regression prediction equations of cartographic image-searching ability as a function of several parameters. Specifically, prediction equations of target location latency and number of correct target locations as a function of display resolution, display focus, target density, and visual angle were developed for map symbols displayed on both black and white and color TV monitors.

Resolution in TV display research is commonly defined as the number of TV raster lines per symbol height. Shurtleff and Owen (1966) used this definition to investigate legibility requirements for alphanumerics and found resolution to influence accuracy and time required to identify symbols. Resolution requirements for other symbols, such as stars, hexagons, rectangles, and circles, were studied by Hemingway and Erickson (1969). Resolution was also studied by Johnston (1969) in a task requiring pilots to locate and identify targets on a terrain model presented on a closed circuit, TV monitor. Horizontal resolution in terms of number of TV raster lines significantly affected the time required for recognition and identification. Preliminary investigations of resolution requirements of cartographic symbols were made by Marsetta and Shurtleff (1966) who used various military unit map symbols. Interestingly, these symbols required a greater number of TV lines for recognition than alphanumerics of the same height. Recently, Wong and Yacoumelos (1970) studied resolution of a closed-circuit, black and white television system used for the identification of topographic symbols. These investigators found resolution to be a function of both TV raster lines per mm of actual map area and the spectral response characteristics of the video system.

In a system in which the observer controls the system equipment, a variable such as focus becomes important. In the course of searching a wide area of topographic material, one might be required to reset focus several times; and, under conditions of environmental stress, focus might become less than perfect. No studies of this variable have been conducted on the TV transmissions of cartographic symbology although Hoffman and Greening (1966) studied a related variable called blur of targets, the poor image quality due to movement across the TV screen.

In a target location task, the factor of density, or amount of nontarget information, is also a determiner of the information processing capabilities of an observer. Baker, Morris, and Steedman (1960) studied this variable in a cathode-ray tube detection task and obtained expected results. Namely, as the number of nontarget objects on the screen increases, search time increases and accuracy decreases. No comparable work, however, has been done with a video task involving search for particular topographic information.

The visual angle of the observer is important in determining his visual acuity. The measure outlined by Morgan, Cook, Chapanis, and Lund (1963) for visual angle is:

$$\text{Visual Angle} = 2 \arctan (d/2D) \qquad (1)$$

where d equals height of the display (or object) and D equals the distance from the observer to the display. A basic visual acuity curve is presented by Morgan, Cook, Chapanis, and Lund (1963) which relates the probability of detection of targets to the visual angle of the target. This curve is important because it is affected by the other parameters of the system as shown in studies by Shurtleff, Marsetta, and Showman (1966) and Baker and Nicholson (1967). Hemingway and Erickson (1969) conducted a similar study and combined their results with the results of the two previous studies. The curves from this combination show that performance is a function of both visual angle of targets on the display and the number of TV raster lines per symbol height.

One limitation of the RSM procedure for investigating the simultaneous effects of these variables is that each variable included in the multiple regression prediction equation is assumed to be quantitative and continuous. Of the variables discussed, nontarget density, focus, and differences between the black and white and the color display may not be quantitative. To include nontarget density in the regression equation it is necessary for it to be semiquantified by defining it in terms of the number of nontarget symbols per map area displayed. Likewise, focus can be arbitrarily quantified by defining it in terms of distance from the plane of sharp image. To investigate the effect of different monitor systems, regression equations predicting performance as a function of display resolution, display focus, visual angle, and target density could be derived separately for the black and white TV monitor and for the color TV monitor. Equal response contours resulting from each prediction equation could then be compared to determine the differential effects of the two TV monitor systems.

Besides illustrating the use of a within-subject, RSM central-composite design, the major purpose of this paper is methodological. Clark and Williges (1972b) discussed two ways of analyzing data collected from a RSM central-composite design in which replication occurs over the complete design. The data could be collapsed across subjects prior to analysis, thereby reducing the design to the traditional RSM central-composite design with repeated observations only at the center; or, alternatively, the collapsed data could be analyzed directly. Both of these analysis procedures were compared in this study in terms of the resulting sensitivity of the analysis and in terms of the predictive validity of the regression equations as determined through cross-validation.

METHOD

Apparatus

The TV system used was a closed-circuit system consisting of a standard 525-line black and white Concord MR-800 monitor, a Setchell Carlson 9MC914 color monitor, and a Sony DXC-5000 color camera. The camera was provided with a VDC-1100 close-up lens with a variable focal length giving the system magnification capability.

Subjects

The subjects who served as observers of the cartographic displays were Army Reserve Officer Training Corps cadets and were familiar with topographic symbology through their course work. These cadets were paid $6.00 for participation in the experiment.

Tasks and Procedures

The observer's task was to locate the position of target symbols on the display monitor. Three point symbols, water towers, schools, and churches, were employed as targets, and the observers were shown examples of these three target types before the session began.

Each experimental condition consisted of a three-trial set. On each trial, a different symbol was used as the target. Within any given set of trials all targets were used but the order of usage was counterbalanced. Each observer sat in front of the monitor and was provided with a long pointer to locate target symbols. The monitor was blanked before each trial began, and the observer was told which symbol was the target for that trial. When the display was revealed, the observer had 60 seconds to locate the target. The three possible outcomes for each trial were: 1) the observer correctly identified the target during the 60-second period, 2) the observer incorrectly pointed to a nontarget symbol, or 3) the observer failed to make a response. In the first case, the time was recorded for detection, and the observer was scored as correct. In the second and third cases, the time recorded was 60 seconds, and the observation was scored as incorrect.

Experimental Design

A four-factor, second-order RSM central-composite design was used (Cochran and Cox, 1957). Basically, the central-composite design consisted of a center point, a 2^K factorial portion, and 2K additional points. Each of the four variables occurred at five levels coded as $-\alpha$, -1, 0, $+1$, $+\alpha$ where ± 1 defined the levels of the factorial portion of the design, $\mp \alpha$ defined additional 2K points, and 0 defined the center point. The design was blocked across days to insure that any differences in testing days would not affect the parameters of the prediction equation. To insure orthogonal blocking, a coded value of α equal to 2 was chosen. (See Clark and Williges, 1972b, for a discussion of the calculation of α.) Table 1 summarizes the coded value coordinates of the data points comprising the design.

- - - - - - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - - - - - -

Because the design was within-subject, each of the six subjects received all 30 treatment conditions shown in Table 1 over a three-day period with one block of 10 treatment conditions presented each day. To minimize the possible differential effects of testing days, block order of presentation was completely counterbalanced across the six subjects. Table 1 shows that the central-composite design was blocked such that one half replicate of the 2^4 factorial design was presented in Block 1, and the other half replicate was presented in Block 2. The fourth-order interaction was chosen as the defining relationship for each half replicate so that no first- or second-order effects would be confounded with blocks or with each other in the second-order RSM central-composite design. Block 3 was composed of the α component of the design. The $\pm\alpha$ value of each variable appeared with only the center (0) value of the other factors. The center point (0, 0, 0, 0) was observed twice in each block in order to obtain an estimate of experimental error.

The four factors included in the design were focus, visual angle, TV raster lines per mm of actual map size, and density. Focus was varied by changing the distance of the TV camera from the plane of sharp image. The levels were 4, 3, 2, 1, and 0 cm from this plane. These values corresponded to linear transformations of the RSM central-composite design coded values of -2, -1, 0, +1, +2, respectively. Visual angle was measured by the arc subtended by the displayed map as determined by Equation 1, and the actual values were 5.00, 6.75, 8.50, 10.25, and 12.00 degrees. TV raster lines per mm was varied by adjusting the focal length of the lens, resulting in real-world values of 4, 5, 6, 7, and 8 TV raster lines. Density was measured by the number of nontarget symbols per map area displayed with actual values of 450, 350, 250, 150, and 50 nontarget symbols. Examples of maps used in this study are shown in Figure 1, which also illustrates the five levels of density and the different target symbols used. Map areas were selected from the 1:24,000 series of United States Geological Survey (USGS) maps of Illinois. To control against learning effects, sufficient maps were collected so that an observer viewed each map only once.

- - - - - - - - - - - - - - - - - -

Insert Figure 1 about here

- - - - - - - - - - - - - - - - - -

RESULTS AND DISCUSSION

The data were analyzed using two different strategies to determine multiple regression equations for prediction as discussed by Clark and Williges (1972b). First, the data were collapsed to produce one score for each treatment condition before analysis. Second, all data for each subject for each treatment condition were analyzed directly. Both analysis strategies were compared and further evaluated in terms of a subsequent cross-validation study. Details on the computer program used to conduct the analyses are discussed by Clark, Williges, and Carmer (1971). Additional details on the mathematical procedures are presented by Clark and Williges (1972a).

Collapsed Median Data Analysis

The data for this analysis were median values across all six subjects on each of the 30 experimental data collection points listed in Table 1. Obtaining a collapsed or median score for each point allowed the data to be analyzed as a standard, blocked, RSM central-composite design. With collapsing, subject effects were eliminated, and experimental error was estimated by the six center points of the RSM central-composite design. The median was chosen as the collapsing statistic so that a markedly different subject would not heavily bias the collapsed score. Calculations of the multiple regression and the subsequent analysis of variance followed the general calculation formulae presented by Williges and Baron (1972).

The major results of these analyses were the multiple regression prediction equations. Separate equations were derived for the black and white monitor and the color monitor. The dependent variables were latency to locate correctly a target and number of correct symbol locations. The resulting first-order prediction equations were:

$$\text{Latency (black and white)} = 38.56 - 2.76F - 6.36D - 0.49V - 1.47T \qquad (2)$$

$$\text{Latency (color)} = 40.04 - 5.54F - 3.60D - 3.71V - 4.08T \qquad (3)$$

$$\text{Correct Locations (black and white)} = 1.76 + 0.21F + 0.34D + 0.09V + 0.12T \qquad (4)$$

$$\text{Correct Locations (color)} = 1.67 + 0.46F + 0.29D + 0.27V + 0.28T \qquad (5)$$

The equations represent the coded values used for F, focus; D, density of nontarget symbols; V, visual angle; and T, TV raster lines per mm of actual map. The respective multiple correlation coefficients were .779, .789, .641, and .848.

Although the weightings of the various parameters differed for the black and white system and the color system, the general effects were consistent. Latency decreased as the coded values of the four predictors increased. The coding was such that as latency decreased, sharp focus, visual angle, and TV lines increased and nontarget density decreased. Similarly, the number of correct target locations increased as the coded values of the various parameters increased.

The reliability of the weightings (partial regression coefficients) of the four parameters of each first-order prediction equation can be tested in an analysis of variance. The various *F* ratios are summarized in Table 2. Focus was a significant predictor in all four equations; however, density was significant only for the black and white system. Visual angle and TV raster lines were not significant ($p > .05$) in any of the collapsed prediction equations.

- - - - - - - - - - - - - - - - -

Insert Table 2 about here

- - - - - - - - - - - - - - - - -

Table 2 also summarizes the *F* tests conducted on blocks and lack of fit. Blocks, as expected, was not significant ($p > .05$) because the order of block presentation over days was completely counterbalanced across the six subjects. Lack of fit was also not significant ($p > .05$). Even though a second-order RSM central-composite design was used for data collection, thereby permitting calculation of a complete second-order equation, the nonsignificant lack of fit suggests that these second-order partial regression coefficients (quadratic effects and linear x linear interactions) may be unreliable predictors if added to the first-order equation.

When the experimenter declares the lack of fit nonsignificant and fails to calculate a higher-order polynomial, he is implicitly accepting the null hypothesis and must consider the probability of declaring an effect nonsignificant when it is actually present. This occurrence, commonly known as a Type II error, can be reduced by increasing the power of the statistical test. One procedure for indirectly increasing power is to choose a higher alpha level or increase the probability of a Type I error. This consideration is noteworthy in connection with results obtained in this study for two of the analyses, namely, number of correct locations and latency on the black and white monitor. If lack of fit were tested at an alpha level of .25, for example, it becomes significant. Fitting a complete second-order equation to both dependent variables of the black and white system as well as both equations for the color monitors, however, yielded no significant second-order partial regression weights. The experimenter, consequently, must decide how much he is willing to trade off a Type I error to reduce a possible Type II error.

Uncollapsed Within-Subject Data Analysis

The second analysis used the data of all six subjects' scores for each experimental condition. The center point (0, 0, 0, 0) of the design represented in Table 1 by observation numbers 9, 10, 19, 20, 29, and 30, was used only once for this analysis. When only one center point is used, the orthogonality of the blocks and treatment effects is not present (Clark and Williges, 1972b). The α length must be changed to accommodate the analysis of blocking effects in this case. This would change the value of the variables for observations 21 - 28. For this analysis, the center point observed first by each subject was used. Because its occurrence fell in different blocks due to counterbalancing and because blocks was not significant in the collapsed analysis, no consideration was given to a blocks effect.

Calculations of the multiple regression followed the same procedure used with the collapsed data although more observations were present. The analysis of variance of the within-subject design required changes in the calculation of error variance. Error variance was obtained from the sum of squares of the replication of the data points (as defined by Williges and Baron, 1972) corrected by subtracting the main effect of subjects. The main effect due to subjects refers to intersubject variability, and this subject variation was calculated using the following general formula:

$$\text{Subject SS} = \sum_{p=1}^{NS} n_S \, (\overline{Y} - \overline{Y}_{S_p})^2 \qquad (6)$$

where $\overline{Y}$ is the grand mean of the dependent variables across all observations; $\overline{Y}_{S_p}$ is the mean of the dependent variables across the observations comprising the p^{th} subject; n_S is the number of times that each subject is observed (a constant value for all subjects); and NS is the number of subjects comprising the entire design. (See Clark and Williges, 1972a, for additional details as to the derivation and calculation of a within-subject RSM central-composite design.)

The resulting first-order, coded, multiple regression prediction equations of target location latency and number of correct symbol locations for each TV monitor were:

$$\text{Latency (black and white)} = 37.60 - 3.32F - 5.28D - 0.52V - 1.39T \qquad (7)$$

$$\text{Latency (color)} = 39.76 - 4.67F - 3.03D - 2.63V - 2.95T \qquad (8)$$

$$\text{Correct Locations (black and white)} = 1.69 + 0.19F + 0.33D + 0.06V + 0.11T \qquad (9)$$

$$\text{Correct Locations (color)} = 1.62 + 0.36F + 0.19D + 0.17V + 0.22T \qquad (10)$$

The respective multiple correlations were .464, .476, .424, and .500.

Although the prediction equations resulting from the uncollapsed analysis were very similar to the prediction equations obtained from the collapsed analysis, the multiple correlations were substantially lower. In other words, the prediction equations accounted for a much smaller percent of total variation when the within-subject variability was included in the uncollapsed design.

Besides retaining the intersubject variability, the within-subject design added more degrees of freedom because replication occurs over the entire design. Increasing the degrees of freedom should result in more sensitive F tests of the partial regression coefficients. The various F ratios resulting from analysis of variance on the four uncollapsed, within-subject regressions are summarized in Table 3. Clearly, more first-order partial regression weights were reliable in the uncollapsed analysis than in the collapsed analysis. In addition, it appears that all four predictors were important in determining performance using the color monitor, whereas focus and density were the primary predictors using the black and white system. Reliable subject differences also occurred under the black and white system, but these effects were completely orthogonal to the prediction equations.

- - - - - - - - - - - - - - - - -

Insert Table 3 about here

- - - - - - - - - - - - - - - - -

The discrepancy between the number of reliable predictors for two TV systems is best explained by examination of the factors contributing to the overall resolution of the two systems. The color image was generated by combining three video signals from red, blue, and green guns; and the picture on the color monitor was a combination of the three pictures produced by these signals. The registration of these pictures was often less than perfect; and, consequently, the overall resolution of that system was somewhat degraded. The black and white monitor, on the other hand, received video signals from the color camera that provided uniform spectral response characteristics which resulted in higher overall system resolution.

TV raster lines per mm of actual map and visual angle were both found to be strong determinants of performance in the studies by Shurtleff (1967) and Baker and Nicholson (1967). The results of this study, however, suggest that the effect of TV raster lines is limited by the overall resolution of the television systems. Wong and Yacoumelos (1970) obtained similar results in that they found overall resolution to be a function of both TV raster lines and spectral response characteristics for color symbols.

Figure 2 presents typical response surfaces that can be obtained from the prediction equations. The axes represent the two significant predictors, focus and density, for the latency score on both the black and white and the color systems as predicted by the uncollapsed regression equations (Equations 7 and 8). Number of TV raster lines was held constant at six, and the visual angle was maintained at eight degrees. The three plotted contours for each monitor system indicate levels of performance in terms of location latency scores of 35, 40, and 45 seconds. These curves illustrate the tradeoffs that must be made between the two independent variables to maintain a given level of latency. By superimposing the contours of the black and white system on the color system, differences in these two nonquantitative variables can be determined. The weightings of focus and density resulted in much steeper slopes on the color system response contours than the surface plotted for the black and white system.

- - - - - - - - - - - - - - - - -

Insert Figure 2 about here

- - - - - - - - - - - - - - - - -

Information presented in terms of these contour plots has important implications for the system designer. If camera focus is to be set and reset during the scanning of topographical information, for example, the system must have the capability of focusing within ranges that will not adversely affect performance. Density of target symbols represents a variable that cannot be easily controlled, because cartographic material varies in density of symbols according to area. But, the results of this study suggest that a nonsystem variable such as density may place restrictions upon the ranges of system variables.

The complete first-order multiple regression analysis performed on the uncollapsed data produced a nonsignificant lack of fit in all cases as shown in Table 3. This suggests that performance was best defined by a linear relationship between the variables, and if higher-order coefficients were used, they might not be reliable. Previous studies, however, have shown that this is not the case for TV raster lines per symbol in alphanumeric recognition. A possible explanation for the nonoccurrence of strong quadratic or higher-order trends may be that the strength of the effects for the other variables, such as focus and density, was great enough to reduce or minimize the higher-order effects of TV raster lines over the range of values used in this study. Care must be taken not to extend the results of this study beyond the range of variables tested.

It is also possible that the experimenter is committing a Type II error when he implicitly accepts the null hypothesis, and he fails to isolate higher-order effects due to TV raster lines. Because the RSM central-composite designs were second-order, an additional complete second-order regression analysis was conducted on the uncollapsed data. No significant second-order effects ($\underline{p} > .05$) occurred for the color monitor. Both regression equations for the black and white system, however, resulted in significant second-order effects. In terms of latency, both the Focus x Focus quadratic effect and the Density x TV Lines linear by linear effect were significant ($\underline{p} < .05$). The Density x TV Lines partial regression coefficient was also significant ($\underline{p} < .05$) for number of correct locations on the black and white monitor. Additional data are necessary to determine whether or not these effects become reliable predictors.

Cross-Validation

From a methodological point of view, cross-validation data served two important purposes in this study. First, these data could be added to the original data to determine if various second-order effects became reliable. Second, and more important, the cross-validation data provided an indication of the predictive validity of the original equations. Specifically, the predictive validities of both first- and second-order prediction equations derived from the collapsed and uncollapsed analyses were compared. A more detailed discussion of the double cross-validation data is presented by North and Williges (1972).

Cross-validation data were obtained by replicating the original design. Care was taken to replicate as closely as possible the design, procedures, equipment, task, and stimulus materials. Six new subjects, who were also Army Reserve Officer Training Corps cadets, were used approximately six months after the original data were collected.

Combining the cross-validation data with the original data resulted in the following uncollapsed, first-order, within-subject, coded regression equations:

$$\text{Latency (black and white)} = 40.09 - 3.42F - 4.65D - 0.88V - 1.88T \qquad (11)$$

$$\text{Latency (color)} = 41.34 - 4.18F - 3.03D - 2.33V - 3.52T \qquad (12)$$

$$\text{Correct Locations (black and white)} = 1.58 + 0.20F + 0.33D + 0.04V + 0.11T \qquad (13)$$

$$\text{Correct Locations (color)} = 1.55 + 0.32F + 0.23D + 0.16V + 0.23T \qquad (14)$$

The respective multiple correlations were .453, .503, .431, and .500.

Even though both the combined, within-subject equations were extremely similar to the original within-subject equations (Equations 7 through 10) and the multiple correlations were virtually the same, the linear effect of TV Lines become a reliable predictor for the black and white monitor in terms of the combined within-subject prediction equations of both the latency and the number of correct locations. The various F ratios for the four combined, first-order prediction equations are presented in Table 4. Note that the additional degrees of freedom gained in the combined data were added primarily to the error term, thereby providing more sensitive F tests. In addition, lack of fit was significant ($p < .05$) for the correct locations prediction equation using the black and white monitor. Results of the complete second-order regression on correct locations demonstrated the Density x TV Lines partial regression weight to be reliable ($p < .05$) using the black and white monitor. This agrees with the results of the less sensitive within-subject analysis of the original data. As discussed earlier, the original within-subject data also suggested possible second-order effects for predictions of latency on the black and white system. Lack of fit was significant at the .10 level in this combined analysis. The complete second-order regression on these data showed both Focus x TV Lines and Density x TV Lines to be significant ($p < .05$). These latter results only partially agree with the original within-subject data analyses. No second-order effects were significant ($p > .05$) in the combined analysis of the color monitor.

- - - - - - - - - - - - - - - - - -

Insert Table 4 about here

- - - - - - - - - - - - - - - - - -

The major results of the cross-validation data analyses were the comparisons of the original multiple correlations to cross-validated multiple correlations to estimate the predictive validity of the equations. The original multiple correlation represents the correlation between the original sample of data (derivation sample) and the scores predicted by the resulting regression equation. The cross-validation multiple correlation is the correlation between the values obtained on the second sample of data (cross-validation sample) and the scores predicted by the original regression equation. Reduction or shrinkage was expected in the cross-validation multiple correlation as compared to the original correlation of this study due mainly to the new sample of subjects and testing times. The obtained cross-validated multiple correlations can be compared to shrinkage of the population multiple correlation as estimated by the modified Wherry formula (Lord and Novick, 1968; Herzberg, 1969) in order to evaluate the relative amount of shrinkage obtained through the collapsed and uncollapsed analyses of the original study.

Table 5 presents the various multiple correlations for the complete first-order prediction equation. When the collapsed prediction equations were used to predict collapsed values in the cross-validation sample, the obtained cross-validation multiple correlation, $R_{\hat{1}2}$, compared favorably with the expected shrinkage, R_S, as shown in the upper portion of Table 5.

- - - - - - - - - - - - - - - - -

Insert Table 5 about here

- - - - - - - - - - - - - - - - -

Generally, a prediction equation is used to predict individual subject performance rather than the average of a particular sample of subjects. This prediction is analogous to predicting uncollapsed data. Using the uncollapsed prediction equations of the original sample to predict these individual scores in the cross-validation, $R_{\hat{1}2}$ compared favorably to R_S as shown in the lower portion of Table 5. On the other hand, the center portion of Table 5 shows that $R_{\hat{1}2}$ was substantially lower than R_S when the collapsed prediction equations were used to predict a new sample of individual subject performance.

Because the original multiple correlation, $R_{\hat{1}1}$, was much higher using the collapsed equations rather than the uncollapsed data, one might be [illegible] into believing that the predictive worth of the collapsed regression equations is better than the uncollapsed equations. These data, however, suggest that the collapsed multiple correlations may grossly overstate the value of the equation if they are used to predict individual subject performance; whereas, multiple correlations from the uncollapsed or within-subject designs provide lower but more realistic estimates of the predictive power of the regression equations.

The analogous multiple correlations were calculated using the complete second-order regression equations rather than the first-order equations. These correlations are presented in Table 6. Essentially, the same shrinkage results occurred in comparing the collapsed versus uncollapsed analyses as presented for the first-order equations. Overall, the original multiple correlations, $R_{\hat{1}1}$, were obviously higher for the second-order equations as compared to the first-order equations because more parameters were used (14 partial regression coefficients in the second-order equation as compared to only 4 in the first-order equation). Because no first-order analyses of the original collapsed and uncollapsed data resulted in significant lack of fit ($p > .05$), the resulting second-order partial regression weights might be unreliable and contribute to greater shrinkage in cross-validation. Indeed, this appears to happen because all but one of the $R_{\hat{1}2}$ values were lower than the predicted shrinkage, R_S, values. Even more striking is the comparison of cross validated multiple correlations, $R_{\hat{1}2}$, of both the first- and second-order regression equations shown in Tables 5 and 6, respectively. In all but one case, the second-order $R_{\hat{1}2}$ values were lower than the corresponding first-order $R_{\hat{1}2}$ values. Consequently, these tenuous second-order effects appear to increase rather than reduce shrinkage.

- - - - - - - - - - - - - - - - - -

Insert Table 6 about here

- - - - - - - - - - - - - - - - - -

These data, then, imply that the more parsimonious approach of selecting the order of the regression equation in accordance with the test of lack of fit provides the more valid and stable overall predication equation. If, on the other hand, the RSM central-composite design is being used for exhaustive search and exploration of a response surface, the experimenter may wisely opt to retain marginally reliable higher-order effects in order to search thoroughly all possible areas of activity in the response surface.

One limitation of the present cross-validation data was the generally low value of the original multiple correlations of the within-subject analyses. Shrinkage on the collapsed data could be somewhat limited by floor effects of these correlations. Nonetheless, all the results are consistently in the predicted direction; consequently, higher $R_{\hat{11}}$ values would probably only further substantiate these results. The low multiple correlations obtained were not altogether unexpected because of procedures used in measuring latency and the small number of targets used in the location task.

CONCLUSIONS

Two general methodological conclusions appear warranted. First, uncollapsed or within-subject analyses as suggested by Clark and Williges (1972b) appear to provide a more sensitive analysis as well as more realistic estimates of the preditive worth of the regression equations as compared to collapsed analyses when predictions of individual performance are made. Second, if the RSM central-composite design is used primarily to provide a general purpose prediction equation, the experimenter may wish to minimize the number of parameters in the prediction equation to minimize storage by determining the order of the prediction equation in accordance with the test of lack of fit.

It is clear from the present results that RSM central-composite design techniques are successful in providing efficient procedures for generating multiple regression prediction equations of variables important in cartographic symbol location tasks. Interestingly, both nonquantitative and quantitative variables can be handled. Nonquantitative variables such as differences between black and white and color monitors must be investigated in terms of separate prediction equations. Focus or density represent variables which can be arbitrarily quantified to be included in the prediction equation. Visual angle and raster lines, on the other hand, represent quantitatively scaled variables that are [illegible] amenable to inclusion in prediction equations.

The results of this study provide [illegible] initial attempt to understand the complex relationship of simultaneous effects of [illegible] variables affecting image interpretability in video systems. Before [illegible] relationships can be determined, additional variables of recognized importance must be considered in the prediction equations. Semple, Heapy, Conway, and Burnette (1971) reviewed several variables of importance in cathode-ray tube displays that were not investigated in this study. Examples of these relevant research parameters mentioned are brightness, contrast ratios, surround illumination, and video bandwidth. Additionally, the capability of RSM to handle nonquantified variables allows study of such items as map type, techniques of cartographic symbol design, and methods for briefing an observer prior to the task. Through the use of the RSM central-composite design, the investigator may now have a method of meaningfully investigating all of these variables.

ACKNOWLEDGMENTS

This research was supported by the Life Sciences Program, Air Force Office of Scientific Research. Dr. Glen Finch, Program Manager, Life Sciences Directorate, was the program monitor. A preliminary version of the collapsed and uncollapsed data was presented in a paper given at the Fifteenth Annual Meeting of the Human Factors Society, New York, New York, October 1971. This study was conducted in cooperation with another research program at the University of Illinois supported by the U. S. Army Engineers Topographic Laboratories. Dr. K. W. Wong served as principal investigator of the second project, and the authors wish to express their appreciation for his suggestions. The help given by N. Yacoumelos throughout the conduct of this study, and the comments given by Beverly H. Williges on earlier versions of this paper were also greatly appreciated.

REFERENCES

Baker, C. A., Morris, D. F., and Steedman, W. C. Target recognition on complex displays. *Human Factors*, 1960, **2**, 51-60.

Baker, C. A. and Nicholson, R. M. Raster scan parameters and target identification. *Proceedings of 19th Annual Aerospace Electronics Conference*, May 1967.

Box, G. E. P. and Wilson, K. B. On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society*, Series B, 1951, **18**, 1-45.

Clark, C. and Williges, R. C. Response surface methodology analyses. Savoy, Ill.: University of Illinois, Institute of Aviation, Aviation Research Laboratory, Technical Report ARL-72-10/AFOSR-72-5, June 1972. (a)

Clark, C. and Williges, R. C. Response surface methodology central-composite design modifications for human performance research. *Human Factors*, 1972, in press. (b)

Clark, C., Williges, R. C., and Carmer, S. General computer program for response surface methodology analyses. Savoy, Ill.: University of Illinois, Institute of Aviation, Aviation Research Laboratory, Technical Report ARL-71-8/AFOSR-71-1, May 1971.

Cochran, W. G. and Cox, G. M. Some methods for the study of response surfaces. *Experimental designs*. New York: Wiley, 1957, 335-375.

Greening, C. P. and Wyman, M. J. Experimental evaluation of a visual detection model. *Human Factors*, 1970, **12**, 435-445.

Hemingway, J. C. and Erickson, R. A. Relative effects of raster scan lines and image subtense on symbol legibility on television. *Human Factors*, 1969, **11**, 331-338.

Herzberg, P. A. The parameters of cross-validation. *Psychometric Monograph Supplement*, 1969, **34**, No. 16.

Hoffman, C. S. and Greening, C. P. The effect of blur and size on target recognition. Paper presented at 37th Annual Scientific Meeting of the Aerospace Medial Association, Las Vegas, Nev., 18-21 April 1966.

Johnston, D. M. Target recognition on TV as a function of horizontal resolution and shades of gray Human Factors, 1969, 10, 201-210.

Lord, F. M. and Novick, M. R. Statistical theories of mental test scores. Chapter 13. The selection prediction variables. Reading, Mass.: Addison-Wesley, 1968, 284-301.

Marsetta, M. and Shurtleff, D. Studies in display symbol legibility. Part XIV: The legibility of military map symbols on television. Bedford, Mass.: The Mitre Corporation, Report MTR 264, September 1966.

Morgan, C. T., Cook, J. S., III, Chapanis, A., and Lund, M. W. Human engineering guide to equipment design. New York: McGraw-Hill, 1963.

North, R. A. and Williges, R. C. Double cross-validation of video cartographic symbol location performance. Proceedings of the Sixteenth Annual Meeting of the Human Factors Society, 1972, in press.

Semple, C. A., Jr., Heapy, R. J., Conway, E. J., Jr., and Burnette, K. T. Analysis of human factors data for electronic flight display systems. Wright-Patterson Air Force Base, Ohio: Air Force Flight Dynamics Laboratory, Technical Report AFFDL-TR-70-174, January 1971.

Shurtleff, D. A. Studies in television legibility: A review of the literature. Information Display, 1967, 4, 40-45.

Shurtleff, D. A., Marsetta, M., and Showman, D. Studies of display symbol legibility: IX. The effects of resolution, size, and viewing angle of legibility. Hanscom Field, Mass.: Electronics Systems Division, Technical Report ESD-TR-65-411, May 1966. (AD 633 833)

Shurtleff, D. A. and Owen, D. Studies of display symbol legibility: VI. A comparison of the legibility of televised leroy and courtney symbols. Hanscom Field, Mass.: Electronic Systems Division, Technical Report ESD-TR-65-136, May 1966. (AD 633 855)

Williges, R. C. and Baron, M. L. Transfer assessment using a between-subjects central-composite design. Human Factors, 1972, in press.

Williges, R. C. and Simon, C. W. Applying response surface methodology to problems of target acquisition. Human Factors, 1971, 13, 511-519.

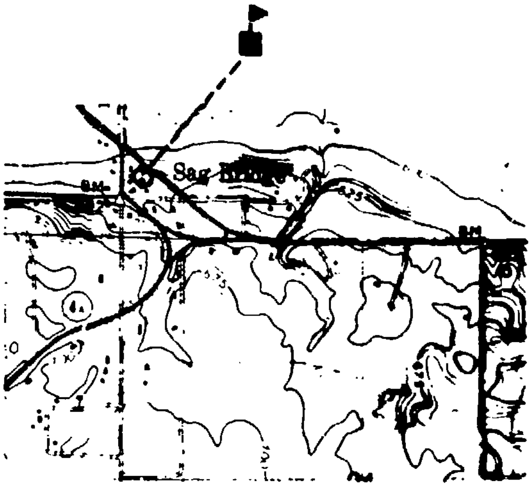
Wong, K. W. and Yacoumelos, N. G. Television display of topographic information. Fort Belvoir, Va.: U. S. Army Engineers Topographic Laboratories, Project 4A061102B52C, Contract DAAK 02-69-C-0628, Contract Report ETL-CR-70-7, November 1970.

TABLE 1

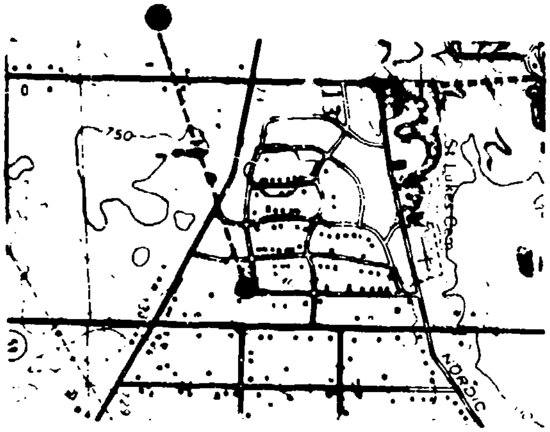
Coded Values for Data Collection Points for Second-Order RSM Central-Composite Design Including Four Variables with Orthogonal Blocking

Treatment Condition	Block	Focus	Density	Visual Angle	TV Raster Lines
1	1	1	1	-1	-1
2	1	1	1	1	1
3	1	1	-1	-1	1
4	1	1	-1	1	-1
5	1	-1	1	1	-1
6	1	-1	1	-1	1
7	1	-1	-1	1	1
8	1	-1	-1	-1	-1
9	1	0	0	0	0
10	1	0	0	0	0
11	2	1	1	1	-1
12	2	1	1	-1	1
13	2	1	-1	1	1
14	2	1	-1	-1	-1
15	2	-1	1	1	1
16	2	-1	1	-1	-1
17	2	-1	-1	1	-1
18	2	-1	-1	-1	1
19	2	0	0	0	0
20	2	0	0	0	0
21	3	0	0	0	2
22	3	0	0	0	-2
23	3	0	0	2	0
24	3	0	0	-2	0
25	3	0	2	0	0
26	3	0	-2	0	0
27	3	2	0	0	0
28	3	-2	0	0	0
29	3	0	0	0	0
30	3	0	0	0	0

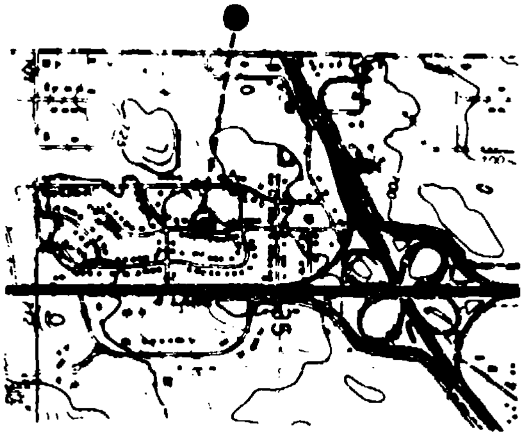
LIST OF FIGURES



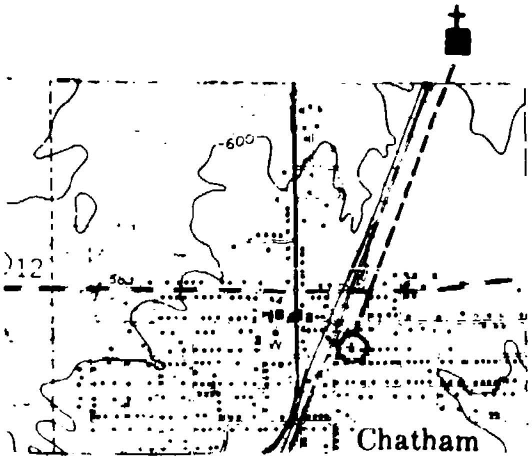
DENSITY = 50



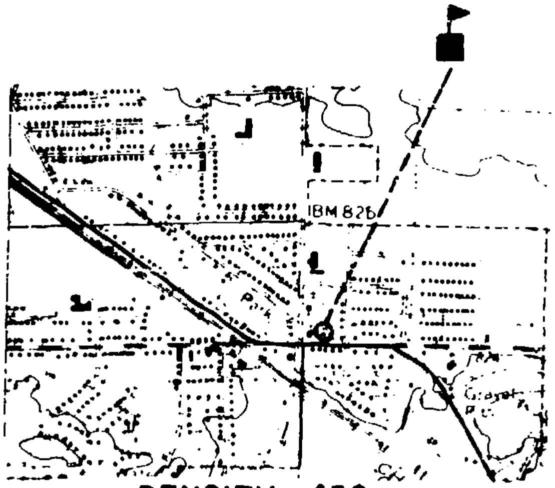
DENSITY = 150



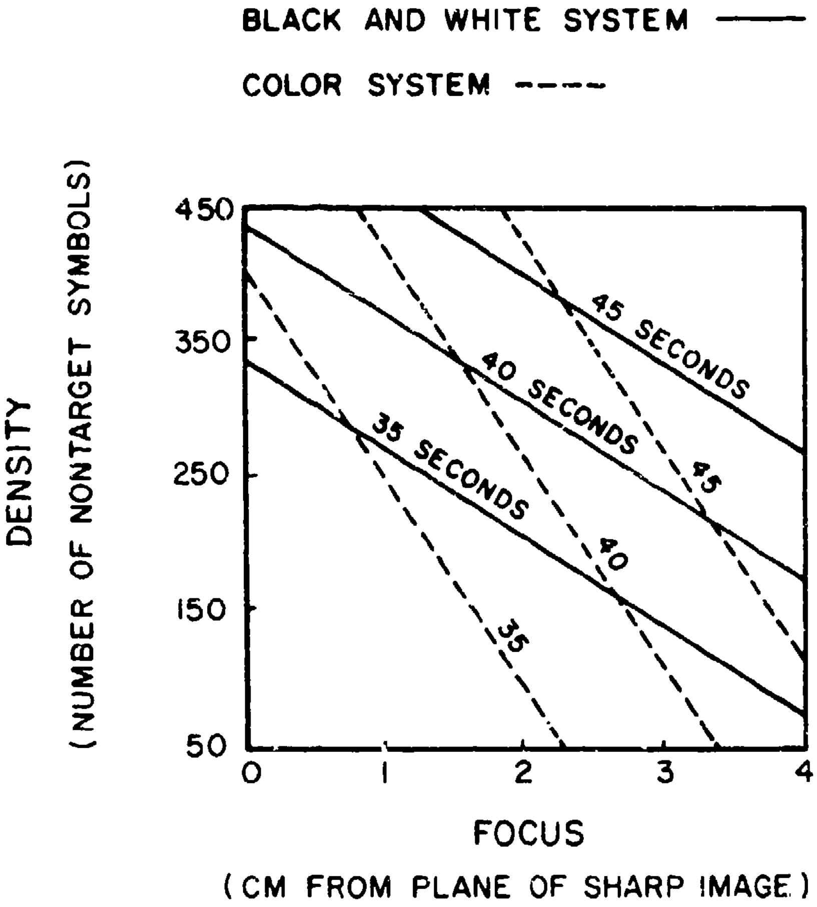
DENSITY = 250



DENSITY = 350



DENSITY = 450


BLACK AND WHITE SYSTEM ——
COLOR SYSTEM ----
DENSITY
(NUMBER OF NONTARGET SYMBOLS)
450
350
250
150
50
0
1
2
3
4
FOCUS
(CM FROM PLANE OF SHARP IMAGE)
45 SECONDS
40 SECONDS
35 SECONDS
45
40
35

BIOGRAPHY

ROBERT C. WILLIGES (Prediction and Cross-Validation of Video Cartographic Symbol Location Performance) is the Assistant Head for Research of the Aviation Research Laboratory of the Institute of Aviation and Associate Professor of Psychology and of Aviation at the University of Illinois. He received his A.B. degree in psychology from Wittenberg University in 1964 and his M.A. and Ph.D. degrees in engineering psychology from The Ohio State University in 1966 and 1968, respectively. While at Ohio State he was a research assistant at the Human Performance Center and conducted research on team training and monitoring of complex computer-generated displays. Prior to joining the Aviation Research Laboratory in 1970, he was Assistant Director of the Highway Traffic Safety Center at the University of Illinois. His current research interests include problems of visual monitoring performance, inspector behavior, and human performance in complex system operation including investigation of rate-field, frequency-separated, and visual time-compressed displays, interpretability of TV-displayed cartographic information, transfer of training, and applications of response surface methodology.

BIOGRAPHY

ROBERT A. NORTH (Prediction and Cross-Validation of Video Cartographic Symbol Location Performance) is a graduate research assistant at the Aviation Research Laboratory at the University of Illinois. He received his B.S. degree in psychology and M.S. degree in engineering psychology from the University of Illinois in 1970 and 1972, respectively. While at Illinois he has performed research in visual monitoring performance as an undergraduate and graduate student. His current interests at the Aviation Research Laboratory include assessing the application of response surface methodology in conjunction with video cartographic symbol location and with variables related to time-compressed radar displays.

Performance Prediction in a Single-Operator Simulated Surveillance System

ROBERT G. MILLS, Aerospace Medical Research Laboratory, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson AFB, Ohio, and ROBERT C. WILLIGES, University of Illinois at Urbana-Champaign

A semiautomatic radar surveillance system was simulated using a time-compressed real-time cathode-ray tube display. Subjects were required to detect targets entering the surveillance area, initiate automatic tracking of these targets, and reinitiate lost tracks when automatic tracking failed. A within-subject Response Surface Methodology (RSM) central-composite design was employed that permitted simultaneous investigation of the effects of five system parameters on surveillance operator performance. Response surface fits (second-order polynomials) were obtained and analyses of variance were conducted to describe these effects on two dependent measures of performance. Results support the contention that operator performance may be dependent upon complex relationships among the five system parameters tested. Furthermore, a RSM central-composite design provided an efficient method for obtaining data and quantifying these relationships.

INTRODUCTION

The main purpose of this report is to present the results of a study of operator capabilities in performing the surveillance tasks of aircraft track initiation and maintenance. The surveillance tasks were performed while monitoring simulated, digitized, and time-compressed radar returns displayed on a computer-graphics display.

Track initiation and maintenance are major functions of present-day semiautomatic air traffic control and surveillance systems such as the Airborne Warning and Control System (AWACS) and new FAA systems presently being developed. Despite the importance of these tasks, however, they have received little attention from human engineering researchers. As a result, human engineering performance criteria important in the design of modern surveillance systems are largely unknown.

Often in these systems radar returns are displayed using time compression of successive radar antenna scans for visual display in real time. Time compression is achieved by storing the digitized returns from successive scans of a radar antenna. These scans are displayed rapidly in proper temporal sequence during the time required to obtain new returns from the next antenna scan, thereby providing the operator with a visual history of scans. As each new scan is stored it is added to the sequence, and the oldest scan is deleted. The effect of this type of display is to generate visible trails for coherent returns such as from a moving aircraft and random points for returns from incoherent sources such as ground, sea, or atmospheric clutter.

The track initiation task requires the operator to initiate automatic tracking of returns potentially belonging to a target. Usually a target is designated with a light pen or cursor,and a switch is activated to initiate a new track. After track initiation an alphanumeric track block is displayed adjacent to each new return from a target.

Track maintenance is required when a target has been lost by the autom atic tracking facility. A failure in automatic tracking is evident to the operator when drifting or misplacement of the alphanumeric track block occurs. Track maintenance is performed in the same manner as track initiation, except that a different switch is used to indicate that the track is old.

A secondary purpose of this report is to provide an example of a rather complex application of a Response Surface Methodology (RSM) central-composite design to th study of human performance. The complexity of the application arises from the fact that the study presented herein is multivariate and investigates the effects of five parameters (factors), each with five levels.

Williges and Simon (1971) indicated that the utility of RSM central-composite designs is that they provide a satisfactory solution to the problem of conducting research studies that are necessarily multivariate and which consist of a large number of parameters and levels ot parameters to be investigated. Typically, a researcher faced with such a study is forced to select a small set of parameters and parameter levels to be investigated using an analysis of variance design.

This was precisely the procedure used in three previous studies of surveillance operator performance (Mills and Bauer, 1971a; 1971b; in press). Each of these studies explored the influence of a limited set of air surveillance system parameters on operat performance. It was recognized rather early, however, that all the parameters under separate investigation were present concurrently in the system and were probably interactive. To evaluate the simultaneous effects of these parameters it was necessary to conduct a multivariate study involving a large set of parameters and parameter values.

These previous studies served as the basis for the present study in that they led to the establishment of a minimum of five parameters which could have a simultaneous influence on operator performance. Thus, it was determined that target introduction rate (number of aircraft entering a surveillance area/unit time) is a powerful factor influencing operator performance. The operational range of introduction rates was also established.

Clutter density (the number of pieces of clutter per square nautical mile or total number per scan) can also influence performance, and the range of parameter values hadbeen established. However, as with introduction rate, the full continuum of effective values of clutter density had not been investigated in a single study.

Target velocity was of particular interest because the data from one of the earlier studies (Mills and Bauer, in press) suggested that performance improves as target velocity is increased to some optimal value. Further increases in target velocity, however, may result in performance degradation. Again, an investigation using a full range of target velocities was necessary in order to establish this relationship.

Two other system parameters not as yet investigated were blip/scan probability (the probability that a target return would be displayed over a series of radar scans) and clutter replacement probability (the probability that a piece of clutter would be replaced by a new piece of clutter on the next scan). Because these parameters can be expressed in terms of probabilities, 0.0 to 1.0, a prior examination of their range was not necessary.

The effective ranges of each of the parameters of interest had been established. However, in no case had the full range of any of these parameters been investigated nor had the combined effects of more than three of the parameters been investigated in a single study.

METHOD

Apparatus

An IBM 2250 cathode-ray tube (CRT) graphics terminal was used for control, display purposes. This terminal had a CRT display surface of 144 square inches (12 x 12 inches). The CRT was coated with P7 phosphor which had a persistence time of 400 milliseconds. The terminal light pen, alphanumeric keyboard, and a programmed function keyboard consisting of 32 response keys were used for operator communication with the computer.

Experimental Display

Figure 1 is an illustration of the CRT display used in this study. The figure is a pictorial representation of a time-exposure photograph taken over six scans, or two radar updates, during an early 20-second period of a mission. A number of targets are shown in Figure 1, several of which have numeric track blocks and, thus, have been initiated. The history of each target trail contained five returns.

Insert Figure 1 about here

The surveillance area simulated was square in shape and represented an actual area of 90,000 square nautical miles. The simulated area was displayed on the CRT in an area of 93.51 square inches and was enclosed by latitude and longitude markings.

Simulated radar returns from targets and clutter were displayed as blue-white, well-focused points. During the persistence period of the phosphor, the points were yellow. The points were approximately 0.01 inch in diameter.

Time compression was accomplished by storing the returns (target and clutter) from each simulated scan of the antenna. During an actual mission simulation, these scans were displayed in real time in a time-compressed mode. The time parameters of display presentation may be found in Mills and Bauer (in press).

Clutter for each scan was distributed statistically according to a combination of uniform and exponential distributions. This method provided a realistic distribution of clutter, unevenly distributed over the surveillance area and containing clumping.

A position error was present in displayed clutter and target returns. Position error simulated the error resulting from signal variations, digitization of analog signals, etc. Target and clutter points were displaced from their true position in X and Y Cartesian coordinates according to a normal distribution with mean error equal to 0 and standard deviation equal to 1 nautical mile.

In effect, position error prohibits the display of a return from the same stationary object from being in exactly the same place in each of a series of scans. As a result, the same clutter point tended to wobble from scan to scan. Returns from a target flying a linear vector were displayed irregularly along the true path of the target.

Tasks

Each subject's tasks were to monitor his surveillance area and to perform the track initiation and maintenance functions. The initiation function required the subject to complete three response actions in any order. These actions were as follows:

1. Use a light pen to indicate the latest displayed return of the set of five returns suspected of representing a target.
2. Input, via the alphanumeric keyboard, the numeric signature (up to three digits) to be assigned to the new track. The numeric input was the integer of the last track initiated increased by the value 1.
3. Press a response key labeled NT (new track).

The maintenance function was performed in the same manner, except that the subject pressed a response key labeled OT (old track) instead of NT. Also, the numeric signature inputed was the signature of the track to be maintained.

A subfunction of the maintenance task was referred to as demand maintenance. On a probabilistic basis (probability of track failure equaled 0.01) a track failure was caused by displaying a track block a random distance from the set of returns belonging to a target. In addition, an asterisk was placed to the left of the signature (see Figure 1). The presence of the asterisk was an indication to the subject to maintain the corresponding track as soon as possible and is analogous to the "trouble track" indication used in certain operational surveillance systems.

A correct maintenance operation restored the track block to its correct coordinate position on the next update. In the case of demand maintenance, the asterisk was also removed. As long as a target remained in the surveillance area, its track block could be restored by a correct maintenance operation. For example, a correct maintenance action has been performed on Track 4 in Figure 1 after the first update shown. The result, as shown in the figure, is a repositioning of the displayed numeral 4 closer to its target on the second update.

When an initiation or maintenance error occurred (for example, attempting to initiate an old track or incorrect track block encoding), an audio signal was immediately returned, indicating that the operation performed had been unacceptable. In the case of correct initiation, the encoded numeric signature track block was automatically assigned and displayed to the right of the latest return of the target.

A counter at the upper right of the screen (see Figure 1) provided the number of the next track to be initiated. Encoded information was displayed at the upper left of the screen as it was inputed.

Figure 1 contains several initiated target tracks with their associated track blocks shown in two updates as a result of the time-exposure representation. For example, Track 15 in the lower left quadrant of Figure 1 has two track blocks of the numeric 15. The upper numeric designates the latest return; the lower numeric is from the previous scan and is visible here only because of the time-exposure format. The number 20 at the upper right of Figure 1 indicates that the next track initiated will be numbered 20. Also shown, is a demand maintenance track, Block 19, and its target trail.

The coordinate position of each track block was updated with each scan, simulating the automatic tracking facility of the computer. Error in this function was simulated by modifying the position of each new track block by a small error term. On the display the track block appeared to have a slight, nonlinear drift in its path (see Figure 1, Track Blocks 8 and 13). If not maintained, the track block would eventually drift out of the surveillance area and disappear. This could occur either before or after the correlated target exited the surveillance area.

Experimental Design

The experimental design employed a five parameter RSM central-composite design. The five parameters were blip/scan ratio (BSR), target introduction rate (TIR), clutter replacement probability (CRP), clutter density (CD), and target velocity (TV). Each parameter had five experimental levels determined by the coded values (-2, -1, 0, 1, 2) according to a second-order, central-composite design as found in Cochran and Cox (1957). The design required 27 experimental observations (missions) per subject.[1]

The actual levels of CD were 20, 50, 80, 110, and 140 pieces of clutter per scan. The actual levels of BSR and CRP were .10, .30, .50, .70, and .90. In the case of BSR a probability of .30, for example, meant that there was a .30 probability that a return from a target would be displayed over a set of scans. The visual effect of a return not being displayed was a larger than usual space between the returns of a target. With a BSR = .10 it is quite possible that the returns from a target would never be displayed and, therefore, could not be initiated.

CRP was the probability that a piece of clutter would be replaced by a new piece of clutter on the next scan. In other words, for CRP = .90, 90 percent of all clutter in a given scan would be in a different position on the next scan. This parameter was included to simulate changes in clutter returns due to changing clutter objects themselves. Variability of CRP was also analogous to changing the signal-to-noise ratio on an operational radar.

The actual levels of TIR were 1.5, 2.25, 3.0, 3.75, and 4.5 targets introduced per minute. Because TIR was a statistically distributed parameter, these are mean values. The standard deviation for each value was set at 1.5 with a range of 0 to 10 targets per minute. A mean TIR value = 2.25 indicates that on the average across scans, 2.25 new aircraft would be introduced into the surveillance area every minute of the mission.

The actual levels of TV were 300, 800, 1300, 1800, and 2300 knots. This parameter was also statistical, and these values are means. The standard deviation selected was 200 knots.

Subjects

Four university seniors made up the subject sample. These subjects had served in a previous study (Mills and Bauer, in press), and each had accumulated at least 54 hours of experience on the tasks to be performed. All subjects were paid volunteers.

Procedure

Subjects completed experimental sessions individually while seated at a computer terminal. During each session, the immediate computer area in which the terminal was located was closed off to all other personnel.

A mission was designed to take 44 minutes of real time. Actual mission times over the simulations varied somewhat due to variations in computer processing requirements during the mission as a function of, for example, number of operator errors. Targets were introduced only during the period of 1 to 40 minutes. Missions were completed at an average of four per week. Only one mission could be completed per day. All performance data were automatically recorded during a mission.

In the first experimental session of an earlier study (Mills and Bauer, in press) subjects had been given written instructions which described (a) the general principles of radar, (b) time compression, (c) the simulation and CRT display, and (d) the initiation and maintenance tasks. After receiving the instructions, subjects had completed a 15-minute practice mission. No additional information or practice was given prior to the start of the present study.

RESULTS AND DISCUSSION

Although a variety of dependent measures were obtained for analysis, for the sake of brevity this discussion will be limited to three of the most important ones.[2] The first is the probability of correct track initiation, P(CI). This variable measured the operator's capability to detect a target and perform the ıctions required for track initiation correctly. The probability was computed by taking the ratio of the number of tracks initiated to the total number of targets introduced in the mission.

The second dependent variable was track initiation time, IT. This variable measured the latency between the time a target was introduced into the operator's surveillance area and the time a track was initiated on it by the operator. Mean IT is the average of these latencies across all initiated tracks in a mission and is, essentially, a measure of the operator's average detection time and the time it takes him to perform all three actions required for correct track initiation.

The third dependent variable of interest was the probability of performing the demand maintenance task correctly. This variable measured the operator's capability to detect and act upon a track failure. The probability was computed by taking the ratio of the number of demand maintenances correctly performed divided by total number of track failures.

All response surface analyses were within-subject analyses of a RSM central-composite design and used a computer program developed by Clark, Williges, and Carmer (1971).

Track Initiation Performance

Table 1 presents the complete second-order response surface fits obtained for P(CI) and mean IT. The multiple regression coefficients for these equations were 0.82 and 0.76 for P(CI) and mean IT, respectively. These equations are the most important results of this study, because they can be used to predict response based upon various engineering design inputs.

- - - - - - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - - - - - -

Overall mean P(CI) across all missions and subjects was 0.67 with standard deviation = 0.24 and range = 0.06 to 1.00. Overall mean IT was 183.58 seconds with standard deviation = 71.94 and range = 46.15 to 362.68 seconds.

Tables 2 and 3 present the regression analyses of variance obtained for the P(CI) and mean IT surfaces, respectively. These tables indicate that the five parameters had a major influence on track initiation performance. The results of the analyses of variance also suggest that the IT response variable was more sensitive to the parameters and their interactions than P(CI). This is not surprising in that P(CI) is primarily a function of absolute detection, whereas IT is a function of both the time to detect a target and the time to perform correct initiation actions.

- -

Insert Tables 2 and 3 about here

- -

Another result of the analysis of variance was that the linear component main effect of target velocity made relatively little contribution to P(CI) response variability. However, the contribution of TV x TV (quadratic component) was significant ($\underline{p} < .05$). In the case of mean IT both linear and quadratic components of target velocity were statistically significant ($\underline{p} < .01$). As will be shown more clearly below, the quadratic effect was the result of an improvement in response as target velocity was increased to a threshold value. Beyond this value, further increases in target velocity no longer yielded response improvement.

The effects of blip/scan ratio and clutter replacement probability were of special interest, because they had not been investigated previously. The analyses in Tables 2 and 3 show that both BSR and CRP linear components were statistically significant ($\underline{p} < .01$) and that BSR was the largest contributor to initiation performance. Furthermore, these parameters were involved in interactions given in Table 3. This observation in conjunction with the fact that the remaining three parameters had previously been shown to affect initiation performance (Mills and Bauer, 1971) demonstrates once again the utility of the RSM central-composite design.

The fact that many interactions did not achieve statistical significance does not necessarily mean that these higher-order terms do not contribute to prediction. The statistical test merely demonstrates that given the particular set of partial regression weights, some of these weights are reliable predictors. The higher-order terms may be correlated; therefore, the individual weightings of these predictors may change if terms are eliminated from the equations. Systematic procedures are needed for eliminating those terms which do not contribute to the multiple regression coefficient.

Tables 2 and 3 also indicate that the overall regression was significant ($p < .01$) as well as the subject effect ($p < .01$). The significant subject effect suggests that there were reliable individual differences between subjects. These differences, however, are orthogonal to the regression and have no effect on the prediction equation.

The significant lack of fit ($p < .01$) obtained for IT in Table 3 suggests that a higher-order fit may be required to develop a more accurate IT response surface. The nonsignificant lack of fit ($p > .05$) for P(CI) in Table 2 suggests that the second-degree fit is adequate. This is further supported by the small F ratio obtained (0.40). In the case of both variables, the lack of fit for linear (first-order) regression was statistically significant ($p < .01$).

Figures 2 and 3 are equal response contour plots for P(CI) and IT, respectively. These plots can aid in interpreting the direction and shape of the functions of the effects indicated in Tables 2 and 3. (The influence of parameter interactions is indicated by the curvilinearity of the contours.)

- -

Insert Figures 2 and 3 about here

- -

The effects of each parameter on P(CI) are presented in Figure 2. Note the change in response as BSR and TIR are varied along the axes. To evaluate the effects of CRP, it is necessary to compare Figure 2a, where CRP = .10, with Figure 2b, where CRP = .90. Although there is an area in Figure 2a where the P(CI) is 1.0, no such area exists in Figure 2b, indicating that the P(CI) was lower when CRP was increased. CD had a similar effect on P(CI). Note the decrease in the area of P(CI) = 1.0 from Figure 2c to Figure 2d and from Figure 2e to Figure 2f. Although the linear effect of TV on P(CI) was statistically nonsignificant ($p > .05$), the pattern of its effect is interesting. A large performance degradation occurred as TV was decreased from 1300 to 300 knots (compare Figure 2a with Figure 2c), but little change in P(CI) occurred when TV was increased from 1300 to 2300 knots (compare Figure 2c with Figure 2e).

Similar comparisons can be made using IT contours. IT response varied with changes in the values of BSR and TV across the axes (see Figures 3a and Figure 3b). In addition, increasing TIR degraded the IT response as shown in Figure 3a and Figure 3b. In Figure 3a the best available IT surface is for IT = 0.0 seconds; whereas in Figure 3b the best surface has increased to IT = 90 seconds. It is also interesting to note that the best surface contours, such as the IT = 0.0 contour in Figure 3a, imply an optimal TV in the area of 1500 knots.

When making these comparisons one should keep in mind that the functions are nonlinear, and their slopes are varying. Thus, interpretation is quite general. The important point is that contours can be obtained using these surface equations for any desired set of engineering values of input parameters.

A thorough examination of contours such as those in Figures 2 and 3 yields a general area of response optimality for P(CI) and IT. The parameter values are TV = 1300; CD = 20; CRP = .5; 1.5 ≦ TIR ≦ 2.7; and .8 ≦ BSR ≦ 1.0. The area of response optimality could conceivably be specified more exactly using partial differentiation of the surface equations. However, the problem is a difficult one requiring that parameters be confined to their experimental ranges. Furthermore, the major purpose of this study was not to seek an optimum response. If the experimenter is interested in systematically determining an optimum, the full range of response s rface methodology procedures, such as method of steepest ascent, should be used. (See Cochran and Cox, 1957, for a more complete discussion.)

Track Maintenance Performance

Examination of the data obtained from the maintenance task, particularly that of demand maintenance, indicates that the subjects tended to drop the task and concentrate on the initiation task. As a result, the obtained response surface equation for the probability of correctly performing demand maintenance yielded a multiple regression coefficient of 0.44. This equation could be expected to account for only 19.36 percent of the variability in response. The overall probability of performing a demand maintenance was 0.29 with standard deviation = .29 and range = 0.0 to 0.96.

The failure of subjects to perform the maintenance task consistently could have resulted from several problems. First, subjects may have found the integration of both initiation and maintenance tasks too difficult in this study. However, it should be remembered that the subjects had considerable experience at the start of the study. It would seem reasonable to expect that they could perform both tasks, at least on the easier missions. Two additional possible explanations are that the instructions failed to emphasize the importance of the maintenance task satisfactorily or that the subjects were not motivated. Regardless of which of these possibilities may have occurred, further investigation of the maintenance task with greater experimental control over subjects is needed.

CONCLUSIONS

This study indicates that surveillance operator performance varies as a function of a complex set of system parameters. To demonstrate this fact and to derive the necessary expressions describing the existing relationships, a RSM central-composite design was used. The utility of this approach was demonstrated in that it provided for efficient data collection, and the observations obtained from the response surface equations do describe complex relationships among the five parameters investigated.

However, further investigation is needed. Subjects failed to integrate the very important maintenance task. This fact most surely will introduce some error in operational generalizability of the response surfaces developed to describe initiation performance, because real operators rarely perform only a single task. In addition, an examination of the predictive validity of these response surface equations is required. Such an examination, if positive, would not only demonstrate the predictive validity of the equations, but also would provide further evidence supporting the utility of the RSM central-composite design approach in developing general purpose prediction equations.

ACKNOWLEDGMENTS

This research was sponsored by the Aerospace Medical Research Laboratory, Aerosapce Medical Division, Air Force Systems Command, Wright-Patterson AFB, Ohio. Additional support for data analyses was provided by the Life Sciences Program, Air Force Office of Scientific Research in connections with a research project conducted by the Aviation Research Laboratory of the Institute of Aviation, University of Illinois at Urbana-Champaign. Dr. Glen Finch, Program Manager, Life Sciences Directorate, was the program monitor for AFOSR. Further reproduction is authorized to satisfy needs of the U. S. Government. This paper is identified as AMRL-TR-72-61. Reprint requests should be sent to Robert G. Mills, AMRL (HES), WPAFB, Ohio 45433.

The authors acknowledge the assistance of Donna Nicolosi, Charles Rinehart, Robert North, and Christine Clark in conducting this study. The authors are also grateful to Captain John Ferrillo (ESD, L. G. Hanscom Field, Bedford, Mass.) for assisting in defining simulation parameters and to Michael A. Bauer for assisting in the development of computer programs.

REFERENCES

Clark, C. and Williges, R. C. Response surface methodology central-composite design modifications for human performance research. Human Factors, 1972, in press.

Clark, C., Williges, R. C., and Carmer, S. G. General computer program for response surface methodology analyses. Savoy, Ill.: University of Illinois, Institute of Aviation, Aviation Research Laboratory, Technical Report ARL-71-8/AFOSR-71-1, May 1971.

Cochran, W. G. and Cox, G. M. Experimental designs. New York: Wiley, 1957.

Mills, R. G. and Bauer, M. A. Aircraft track initation and maintenance in a single-operator simulated surveillance system: Technical Report I. Wright-Patterson Air Force Base, Ohio: Aerospace Medical Research Laboratory, Technical Report AMRL-TR-70-103, 1971. (a)

Mills, R. G. and Bauer, M. A. Aircraft track initiation and maintenance in a single-operator simulated surveillance system: Technical Report II. Wright-Patterson Air Force Base, Ohio: Aerospace Medical Research Laboratory, Technical Report AMRL-TR-71-76, 1971. (b)

Mills, R. G. and Bauer, M. A. Aircraft track initiation and maintenance in a single-operator simulated surveillance system: Technical Report III. Wright-Patterson Air Force Base, Ohio: Aerospace Medical Research Laboratory, in press.

Williges, R. C. and Simon, C. W. Applying response surface methodology to problems of target acquisition. Human Factors, 1971, 13, 511-519.

FOOTNOTES

[1] The observation (0, 0, 0, 0, 0) was used only once in the analyses as suggested by Clark and Williges (1972).

[2] A complete presentation of the results obtained for all dependent variables measured will be available in a later Aerospace Medical Research Laboratory Technical Report, in press.

TABLE 1

Second-Order Multiple Regression Prediction Equations for Probability Correct Track Initiation, P(CI) and Mean Track Initiation Time, IT

$$
\begin{aligned}
P(CI) = {} & .293 + 2.193\ BSR - .023\ TIR - .303\ CRP - .002\ CD \\
& + .0009\ TV - 1.285\ BSR \times BSR - .128\ BSR \times TIR + .290\ BSR \times CRP \\
& + .0002\ BSR \times CD + .0002\ BSR \times TV - .004\ TIR \times TIR \\
& + .032\ TIR \times CRP - .0002\ TIR \times CD + .00003\ TIR \times TV \\
& - .090\ CRP \times CRP + .0002\ CRP \times CD - .00002\ CRP \times TV \\
& + .00001\ CD \times CD - .0000004\ CD \times TV - .000001\ TV \times TV
\end{aligned}
$$

$$
\begin{aligned}
IT = {} & 409.18 - 237.20\ BSR - 1.34\ TIR + 128.80\ CRP - .80\ CD \\
& - .23\ TV - 221.59\ BSR \times BSR + 59.77\ BSR \times TIR - 195.68\ BSR \times CRP \\
& + .80\ BSR \times CD + .09\ BSR \times TV - 2.31\ TIR \times TIR \\
& + 18.78\ TIR \times CRP + .22\ TIR \times CD + .002\ TIR \times TV \\
& + 102.36\ CRP \times CRP - .03\ CRP \times CD - .01\ CRP \times TV \\
& - .001\ CD \times CD + .0002\ CD \times TV + .00005\ TV \times TV
\end{aligned}
$$

where	BSR	=	blip/scan ratio
	TIR	=	target introduction rate
	CRP	=	clutter replacement probability
	CD	=	clutter density
	TV	=	target velocity

TABLE 2

Second-Order Regression Analysis of Variance Summary Table for Probability of Correct Track Initiation

Source	df	MS	F
Regression	(20)	2.13×10^{-1}	22.74**
Blip/Scan Ratio (BSR)	1	3.34	356.81**
Target Introduction Rate (TIR)	1	2.72×10^{-1}	29.05**
Clutter Replacement Probability (CRP)	1	9.69×10^{-2}	10.35**
Clutter Density (CD)	1	1.45×10^{-1}	15.48**
Target Velocity (TV)	1	3.01×10^{-4}	0.03
BSR x BSR	1	1.69×10^{-1}	18.06**
BSR x TIR	1	2.36×10^{-2}	2.52
BSR x CRP	1	1.47×10^{-2}	1.57
BSR x CD	1	7.66×10^{-5}	0.01
BSR x TV	1	3.47×10^{-2}	3.70
TIR x TIR	1	2.64×10^{-4}	0.03
TIR x CRP	1	1.50×10^{-3}	0.16
TIR x CD	1	1.91×10^{-3}	0.02
TIR x TV	1	8.79×10^{-3}	0.94
CRP x CRP	1	8.27×10^{-4}	0.09
CRP x CD	1	1.27×10^{-4}	0.01
CRP x TV	1	2.64×10^{-4}	0.03
CD x CD	1	5.44×10^{-3}	0.58
CD x TV	1	1.90×10^{-3}	0.20
TV x TV	1	3.85×10^{-2}	4.11*
Residual	(87)	2.41×10^{-2}	
Subjects	3	4.49×10^{-1}	47.92**
Lack of Fit	6	3.73×10^{-3}	0.40
Replications [a]	78	9.36×10^{-3}	
Total	(107)		

[a] Error term used in F tests

* $p < .05$

** $p < .01$

TABLE 3

Second-Order Regression Analysis of Variance Summary Table for Mean Track Initiation Time

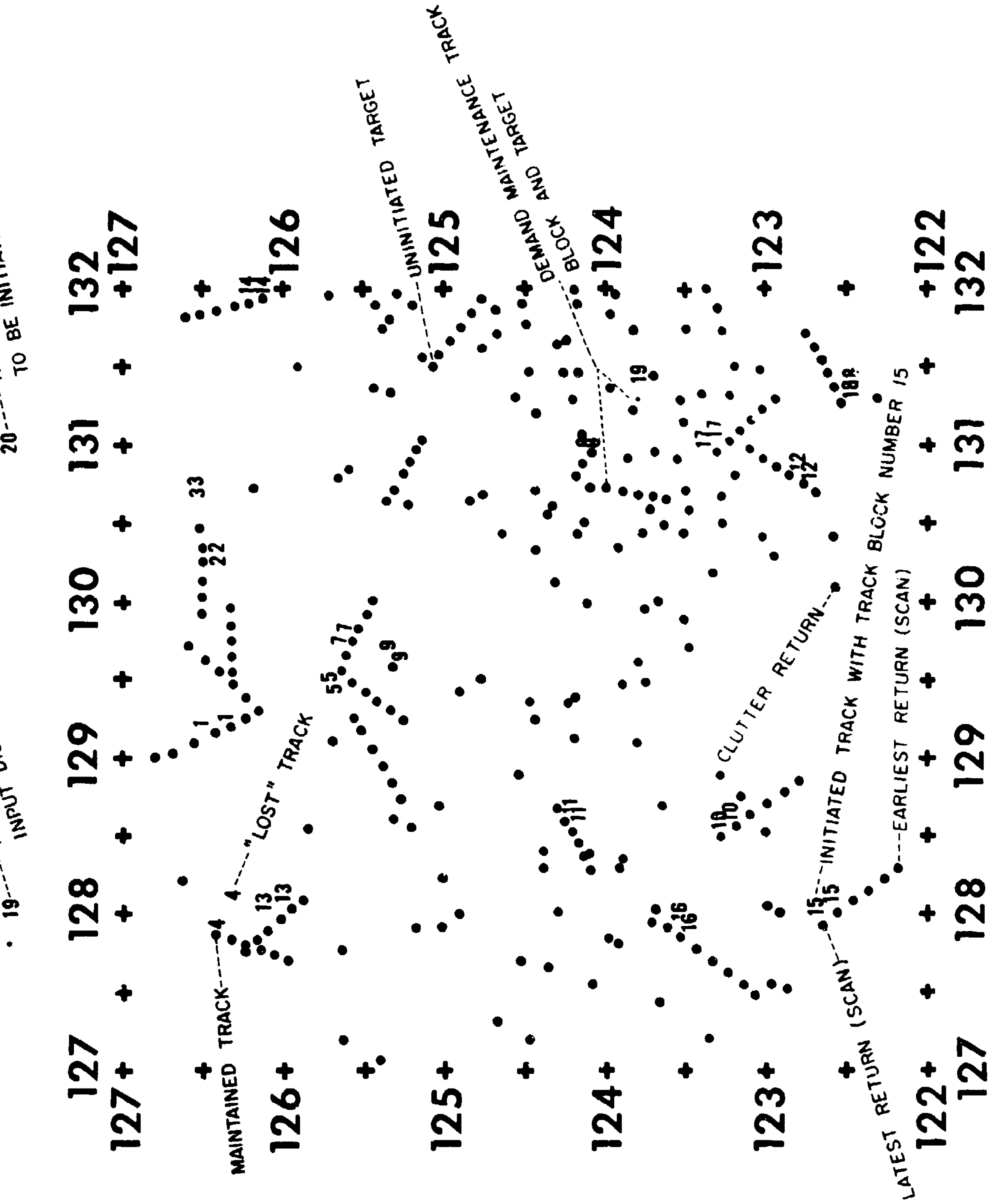
Source	df	MS	F
Regression	(20)	16128.37	16.51**
Blip/Scan Ratio (BSR)	1	144726.40	148.11**
Target Introduction Rate (TIR)	1	33783.38	34.57**
Clutter Replacement Probability (CRP)	1	12397.62	12.69**
Clutter Density (CD)	1	12206.09	12.49**
Target Velocity (TV)	1	72349.97	74.04**
BSR x BSR	1	5028.17	5.15*
BSR x TIR	1	5143.94	5.26*
BSR x CRP	1	3921.11	4.01*
BSR x CD	1	1457.04	1.49
BSR x TV	1	5414.57	5.54*
TIR x TIR	1	107.79	0.11
TIR x CRP	1	507.88	0.52
TIR x CD	1	1516.62	1.55
TIR x TV	1	30.54	0.03
CRP x CRP	1	1072.92	1.10
CRP x CD	1	1.92	0.002
CRP x TV	1	120.64	0.12
CD x CD	1	18.90	0.02
CD x TV	1	494.45	0.51
TV x TV	1	8260.61	8.45**
Residual	(87)	2716.26	
Subjects	3	38902.58	39.81**
Lack of Fit	6	7231.68	7.40**
Replications [a]	78	977.14	
Total	(107)		

[a] Error term used in F tests

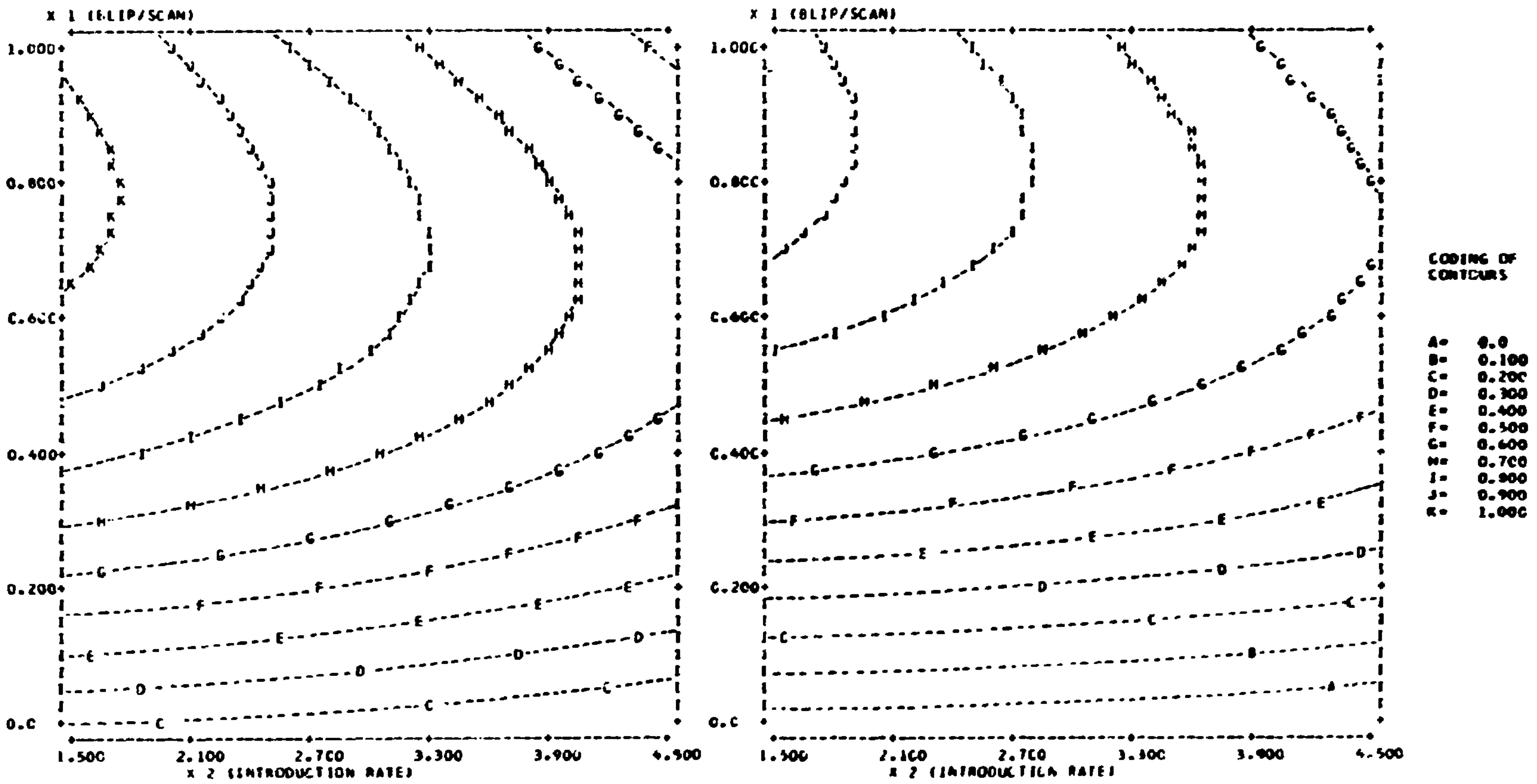
* $p < .05$

** $p < .01$

LIST OF FIGURES



19---TRACK BLOCK INPUT DISPLAY
20---NUMBER OF NEXT TRACK TO BE INITIATED
UNINITIATED TARGET
DEMAND MAINTENANCE TRACK
BLOCK AND TARGET
MAINTAINED TRACK
"LOST" TRACK
CLUTTER RETURN
INITIATED TRACK WITH TRACK BLOCK NUMBER 15
EARLIEST RETURN (SCAN)
LATEST RETURN (SCAN)
127
128
129
130
131
132
127
126
125
124
123
122

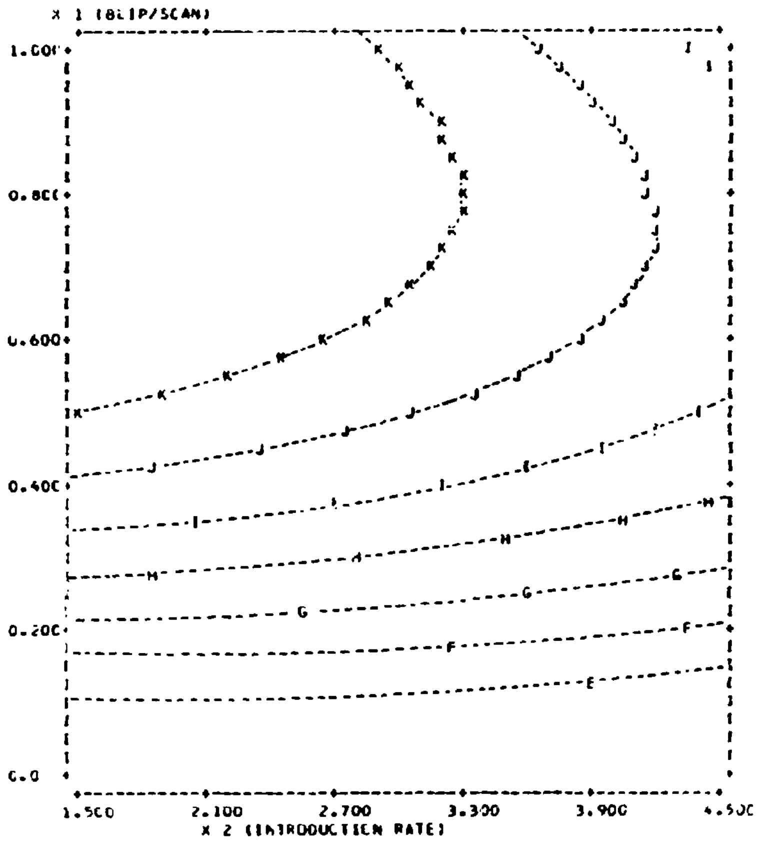


(a)

X 3= 0.100 (CLUTTER REPLACEMENT)
X 4= 20.000 (CLUTTER DENSITY)
X 5= 300.000 (VELOCITY)

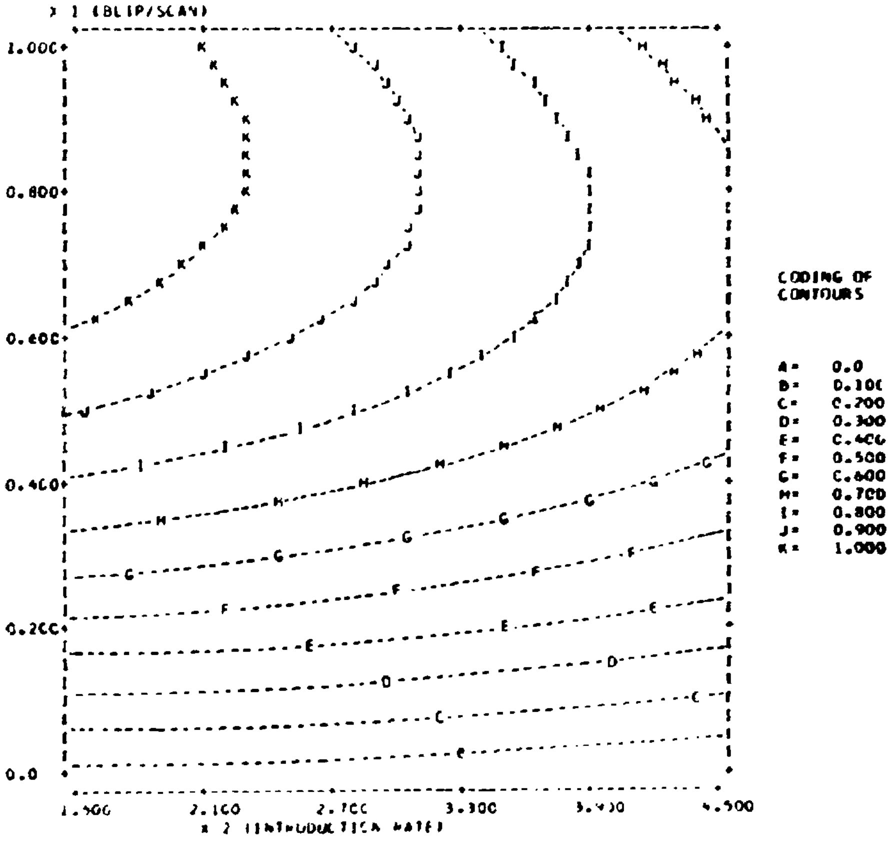
(b)

X 3= 0.900 (CLUTTER REPLACEMENT)
X 4= 20.000 (CLUTTER DENSITY)
X 5= 300.000 (VELOCITY)



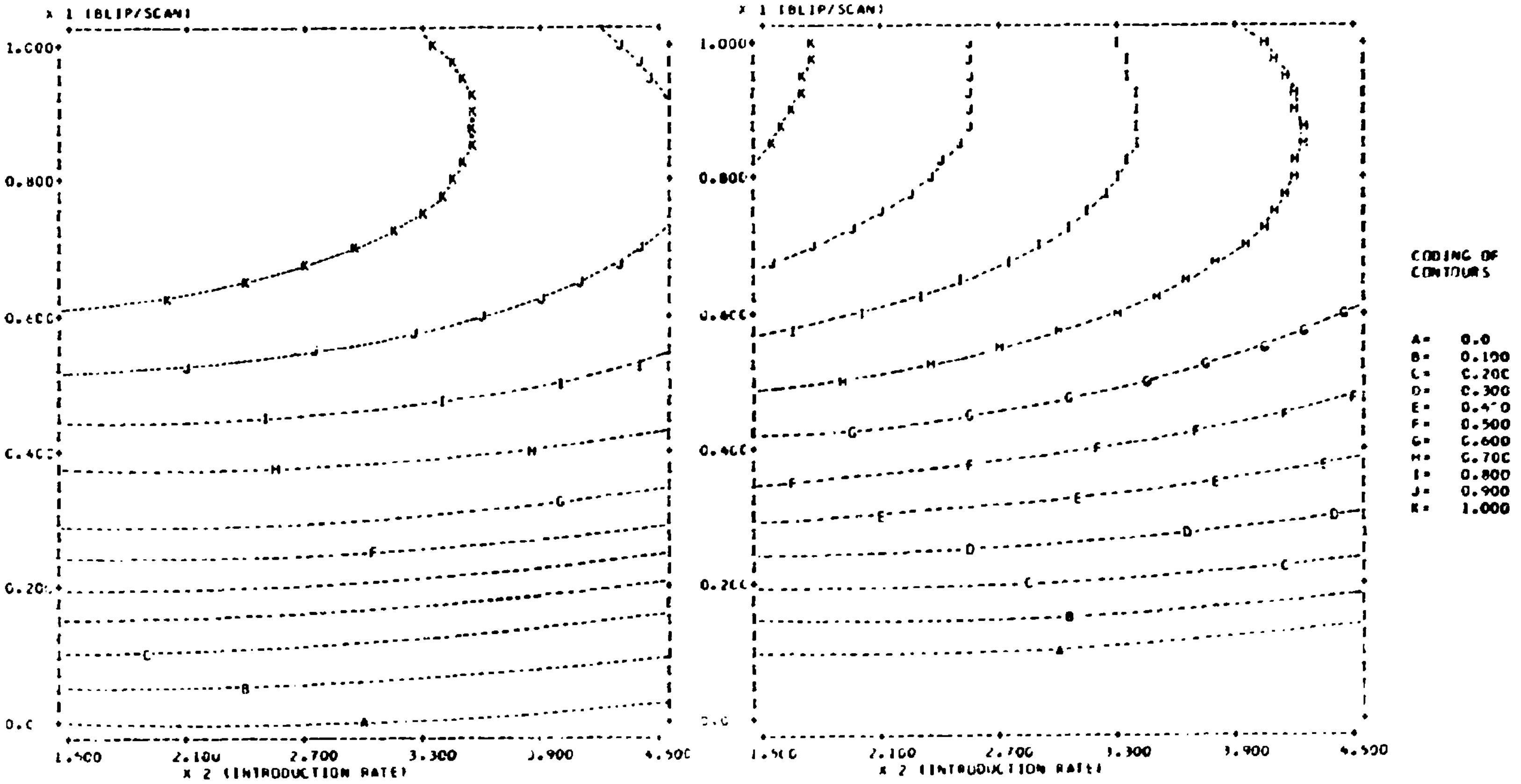
(c)

X 3= 0.1CC (CLUTTER REPLACEMENT)
X 4= 20.000 (CLUTTER DENSITY)
X 5=13C0.COC (VELOCITY)



(d)

X 3= 0.1C0 (CLUTTER REPLACEMENT)
X 4= 80.CCO (CLUTTER DENSITY)
X 5=13CC.COD (VELOCITY)

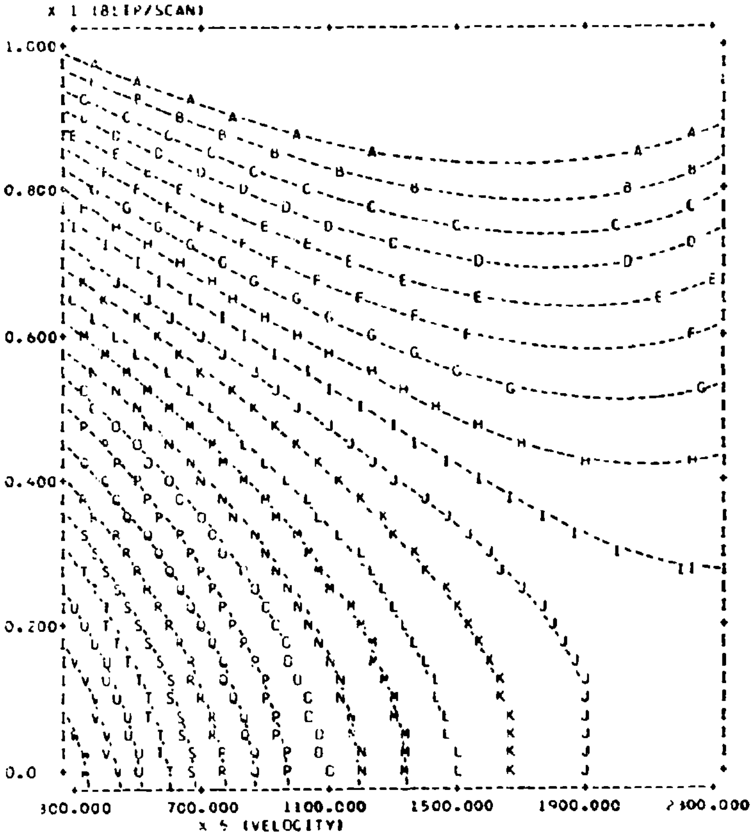


(e)

X 3= 0.100 (CLUTTER REPLACEMENT)
X 4= 20.000 (CLUTTER DENSITY)
X 5=2300.000 (VELOCITY)

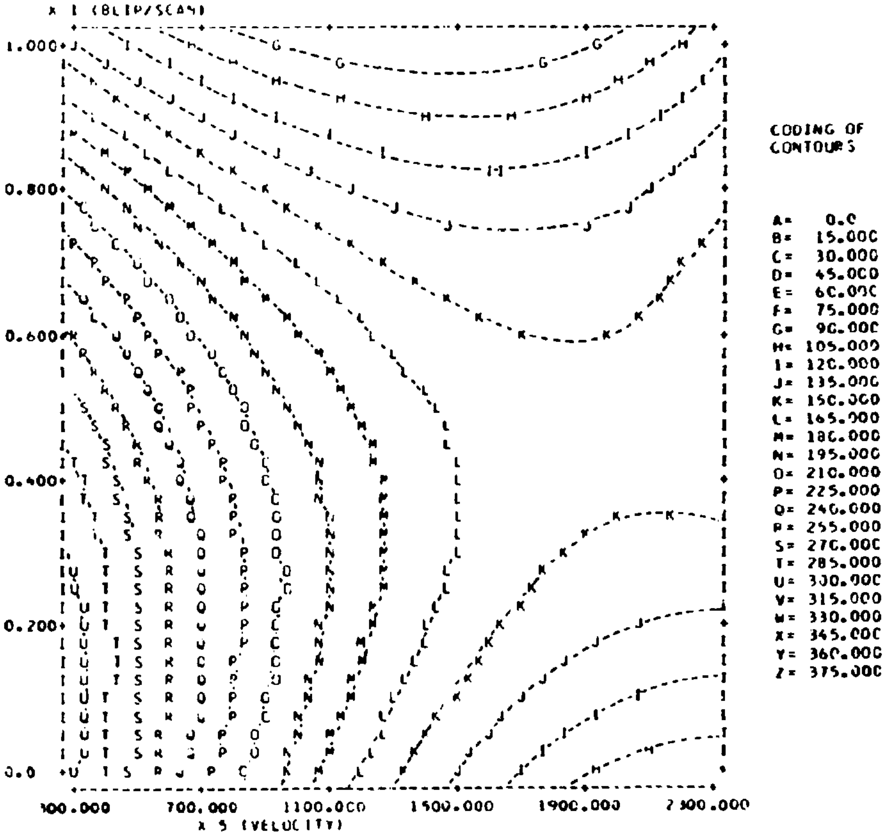
(f)

X 3= 0.100 (CLUTTER REPLACEMENT)
X 4= 140.000 (CLUTTER DENSITY)
X 5=2300.000 (VELOCITY)



X 2= 1.500 (INTRODUCTION RATE)
X 3= 0.100 (CLUTTER REPLACEMENT)
X 4= 20.000 (CLUTTER DENSITY)

(a)



X 2= 4.500 (INTRODUCTION RATE)
X 3= 0.100 (CLUTTER REPLACEMENT)
X 4= 20.000 (CLUTTER DENSITY)

(b)

BIOGRAPHY

ROBERT G. MILLS (Performance Prediction in a Single-Operator Simulated Surveillance System) is an engineering psychologist in the Systems Effectiveness Branch, Human Engineering Division of the Aerospace Medical Research Laboratory, Wright-Patterson AFB, Ohio. He received his B.A. and M.A. degrees from Southern Illinois University in 1961 and 1964, respectively. He is presently working toward a Ph.D. degree in industrial engineering from The Ohio State University, under government sponsorship. His areas of interest include man-machine sysrem effectiveness, computer simulation and modeling, and decision making.

BIOGRAPHY

ROBERT C. WILLIGES (Performance Prediction in a Single-Operator Surveillance System) is the Assistant Head for Research of the Aviation Research Laboratory of the Institute of Aviation and Associate Professor of Psychology and of Aviation at the University of Illinois. He received his A.B. degree in psychology from Wittenberg University in 1964 and his M.A. and Ph.D. degrees in engineering psychology from The Ohio State University in 1966 and 1968, respectively. While at Ohio State he was a research assistant at the Human Performance Center and conducted research on team training and monitoring of complex computer-generated displays. Prior to joining the Aviation Research Laboratory in 1970, he was Assistant Director of the Highway Traffic Safety Center at the University of Illinois. His current research interests include problems of visual monitoring performance, inspector behavior, and human performance in complex system operation including investigation of rate-field, frequency-separated, and visual time-compressed displays, interpretability of TV-displayed cartographic information, transfer of training, and applications of response surface methodology.

Predictive Validity of Central-Composite Design Regression Equations

ROBERT C. WILLIGES, University of Illinois at Urbana-Champaign, and
ROBERT G. MILLS, Aerospace Medical Research Laboratory, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson AFB, Ohio

The predictive validity of the Mills and Williges (1972) empirically derived prediction equations of single operator performance in a simulated surveillance system was assessed by measuring 16 additional data points on the same four subjects participating in the original study. Correlations between predicted and observed performance on 16 points augmented to the design compared favorably with estimated shrunken multiple correlation coefficients. In addition, the averages of each of the 16 additional treatment conditions were compared to the 95 percent confidence interval of the predicted values using the Mills and Williges (1972) regression equations. The 16 data points were also chosen such that a supplementary factorial analysis of variance could be conducted on the data. Comparisons were made between the analysis of variance and the multiple regression analysis. It was concluded that the Response Surface Methodology procedures for developing overall prediction equations of human performance demonstrate a high degree of predictive validity.

INTRODUCTION

A Response Surface Methodology (RSM) central-composite design was used by Mills and Williges (1972) to develop generalized prediction equations for probability of correct initiation and track initiation latency in a simulated surveillance system. For example, track initiation performance was predicted by a second-order multiple regression equation. One primary consideration in assessing the utility of such an empirically derived prediction equation is predictive validity. Shrinkage of the multiple regression coefficient can be expected when the prediction equation developed on one set of subjects is used to predict performance on a new set of subjects. Generally, it is advisable to cross-validate the prediction equation before using it or to estimate the amount of shrinkage in terms of the modified Wherry procedure (Lord and Novick, 1968; and Herzberg, 1969).

Williges and North (1972) demonstrated that a within-subject multiple regression prediction equation of video cartographic image interpretability derived from a RSM central-composite design maintained a multiple correlation with only slight shrinkage under cross-validation to a new set of subjects. The purpose of the present study was to investigate the predictive validity of the RSM regression equation from another point of view.

When a single RSM design is used to predict a fairly large surface, the data points are sparsely distributed across the region of experimental interest. Conceivably, much of the orderly relationship among sampled experimental points of the response surface could be overlooked. The present study compared observed performance at data points not originally sampled in the Mills and Williges (1972) study to the performance predicted by the empirical regression equation of that study in order to assess the predictive validity of the RSM procedure for other points within the surface. In addition, the additional data points were chosen such that a conventional analysis of variance could be conducted on the resulting two-level factorial design without any main effects or interactions confounded.

METHOD

Subjects

To minimize shrinkage due to subject differences, the same four subjects used in the Mills and Williges (1972) study participated in this experiment. Each subject was paid for h:; participation.

Task and Procedures

The experimental task and procedures were identical to those used by Mills and Williges (1972). Data were collected immediately following the completion of that study. The reader is referred to the original study for details of the simulated surveillance system task and the specific experimental procedures.

Design

Coded values of the 27 treatment conditions used in the Mills and Williges (1972) study are listed in Table 1. Note that the first 16 data points represent a one-half fractional replicate of a 2^5 factorial design of the five factors, blip/scan ratio, target introduction rate, clutter replacement probability, clutter density, and target velocity. Coded values of the 16 additional data points used in this study are presented in Table 2. The recoded values were merely linear tran.formation of the real-world values of the various levels of the five factors provided by Mills and Williges (1972). The additional data points were chosen such that the first 16 points originally investigated (see Table 1) combined with the treatment conditions of this study would provide a complete 2^5 factorial design of the five factors.

- - - - - - - - - - - - - - - - - - -

Insert Table 1 and Table 2 about here

- - - - - - - - - - - - - - - - - - -

RESULTS AND DISCUSSION

Regression Analysis

An estimate of the predictive worth of a multiple regression equation can be determined from the multiple regression coefficient, which is the correlation between the observed values of the data and the predicted values obtained from the regression equation. The square of the multiple regression coefficient, the coefficient of determination, indicates the percent of variation accounted for by the regression equation. By correlating the observed responses at the 16 additional data points with the predicted values at these points using the Mills and Williges (1972) regression equations, the resulting correlation coefficient provided an indication of the predictive validity of the regression equations. In addition, this correlation can also be compared to an estimate of the amount of expected shrinkage of the original multiple correlation coefficient. If the equation has high predictive validity, the multiple correlation coefficient should compare favorably with the estimated shrinkage. The shrunken multiple correlation used as a comparative baseline for these data was determined by the modified Wherry formula (Lord and Novick, 1968, and Herzberg, 1969):

$$R_S = \sqrt{1 - (1 - R^2_{\hat{1}1}) \frac{N-1}{N-p-1}} \quad , \qquad (1)$$

where N equals the number of observations used to determine the multiple regression equation and p is the number of partial regression weights or parameters of the multiple regression prediction equation.

Table 3 presents correlations for both the probability of correct initiation and the mean initiation latency in terms of the original Mills and Williges (1972) multiple correlations, $R_{\hat{1}1}$, the shrunken multiple correlations, R_S, and the correlation between the predicted scores and the obtained scores from this study, $R_{\hat{1}2}$. It is obvious from the comparison of the values of $R_{\hat{1}2}$ and R_S that the correlation between the predicted values and the values of the 16 observed data points for each of the four subjects was essentially the same as the predicted shrinkage. Clearly, these data suggest a high predictive validity of the empirical regression equations.

- - - - - - - - - - - - - - - - - -

Insert Table 3 about here

- - - - - - - - - - - - - - - - - -

Another means of assessing the predictive worth of the regression equation is to compare the average of the 16 observed responses across the four subjects to the confidence interval of the predicted values of the Mills and Williges (1972) regression equations. According to Li (1964) the confidence interval of the adjusted mean can be constructed using a *t* distribution and a standard error equal to:

$$\sigma_{\overline{Y}_X} = \sqrt{\frac{\sigma^2}{W}} \qquad (2)$$

where σ^2 is the replication mean square and $1/W = [X_i]\,[c_{ij}]\,[X_i]$ such that $[X_i]$ is the transpose or row vector of the particular levels of the various X values, $[c_{ij}]$ is the inverse of the m + 1 by m + 1 uncorrected sum of squares cross-product matrix, and $[X_i]$ is the column vector of the particular levels of the various X values. Note that the standard error changes according to the particular X values chosen. Because the 16 additional data points used in this study were equidistant from the center (each consisted of various coded combinations of +1 or -1), each of these data points has the same standard error. Using Equation 2, the standard error of the adjusted mean was 0.045 and 14.57 for the probability of correct initiation and mean initiation latency, respectively.

A comparison of the mean observed values on the 16 additional treatment conditions with the 95 percent confidence interval of the Mills and Williges (1972) prediction equations is presented in Table 4. In terms of the probability of correct initiation, all of the obtained probabilities fell within the 95 percent confidence limit of the prediction equation. On the other hand, only five values of mean observed target initation latency fell beyond these confidence limits. These results are certainly compatible with the multiple correlations which suggest that the probability of correct initiation yielded a slightly better prediction equation than the mean initiation latency equation. Both equations, however, appeared to provide relatively accurate and stable predictions.

- - - - - - - - - - - - - - - - -

Insert Table 4 about here

- - - - - - - - - - - - - - - - -

The results of this study are limited to data falling within the range of values of the originally sampled data. If one attempted to predict beyond the ±2 coded value of any factor, the predictive validity could drop markedly, because no attempt was made to measure such trends in the original central-composite design. If, on the other hand, prediction is restricted to within the ±2 coded value, these data support the contention that the predictive validity is high.

Analysis of Variance

Because the additional 16 data points of this study were chosen to complete a factorial design, a 2^5 within-subject analysis of variance could be conducted on both the probability of correct initiation and the mean detection latency. The significant effects for both the analysis of variance of probability of correct initiation and mean initiation latency are summarized in Table 5.

- - - - - - - - - - - - - - - - -

Insert Table 5 about here

- - - - - - - - - - - - - - - - -

Two major difficulties arise when one attempts to compare the results of the analysis of variance with the multiple regression analysis. First, each analysis was addressed to somewhat different experimental questions. The regression equation was directed toward determining a functional relationship among various independent variables and establishing which of these combinations of independent variables were reliable in predicting performance on the dependent variables. Analysis of variance, on the other hand, was addressed to a specific yes-no question; namely, was performance as measured by a dependent variable reliably different when observed at different levels of various individual independent variables (main effects) or at specific levels of certain combinations of independent variables (interactions).

The second difficulty in comparing the results of the two procedures was that different data sets were used in the two analyses. The Mills and Williges (1972) regression analyses were based on a RSM central-composite design that measured performance at selected treatment combinations across five levels of each independent variable; whereas, the analysis of variance included data from only two levels of each independent variable. Consequently, reliable trends appearing in the regression analysis might be occurring primarily beyond the levels measured in the analysis of variance. In addition, the second-order regression equations provided by Mills and Williges (1972) included certain quadratic terms that could not be measured in the analysis of variance design because only two levels were used. On the other hand, the analysis of variance demonstrated certain significant third- and fourth-order linear interactions that could not appear in the second-order regression equations.

Where comparisons could be made between the two analyses, the results were consistent. Both analyses included linear main effects and linear by linear two-way interactions. All of these significant effects resulting from the analysis of variance were also significant predictors in the Mills and Williges (1972) prediction equations. Moreover, the direction of the effects was the same. For example, as blip/scan ratio increased, its linear component significantly increased the probability of correct initiation and decreased the mean latency of track initiation according to the Mills and Williges (1972) prediction equations. Likewise, the significant main effect of blip/scan ratio in the present analysis of variance demonstrated a higher probability of correct initiation and a lower mean latency of target detection as blip/scan ratio increased from the -l level to the +1 level. If the intention of the experimenter is to predict functional relationships, the regression equation is more useful than the traditional analysis of variance even though the results were compatible.

CONCLUSIONS

It appears that adding points to complex RSM central-composite designs so that a 2^K factorial design exists is a useful procedure for assessing the predictive validity of the multiple regression prediction equations as well as allowing calculation of a supplementary factorial analysis of variance on the data. The measure of predictive validity obtained from this study by correlating observed performance on the 16 additional data points with the predicted performance and the results of the cross-validation data provided by Williges and North (1972) provide support for the contention that the RSM central-composite design is an efficient way to generate relatively stable and valid prediction equations of human performance.

ACKNOWLEDGMENTS

This research was sponsored by the Aerospace Medical Research Laboratory, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson AFB, Ohio. Additional support for data analyses was provided by the Life Sciences Program, Air Force Office of Scientific Research in connection with a research project conducted by the Aviation Research Laboratory of the Institute of Aviation, University of Illinois at Urbana-Champaign. Dr. Glen Finch, Program Manager, Life Sciences Directorate, was the program monitor for AFOSR. This paper is identified as AMRL-TR-72-63. Further reproduction is authorized to satisfy needs of the U. S. Government.

The authors acknowledge the assistance of Donna Nicolosi, Charles Rinehart, Robert North, and Chrinstine Clark in conducting this study. The authors are also grateful to Captain John Ferrillo (ESD, L. G. Hanscom Field, Bedford, Mass.) for assisting in defining simulation parameters and to Michael A. Bauer for assisting in the development of computer programs.

REFERENCES

Herzberg, P. A. The parameters of cross-validation. Psychometric Monographs Supplement, 1969, 34, No. 16.

Li, J. C. R. Statistical inference II The multiple regression and its ramifications. Ann Arbor: Edwards Brothers, 1964.

Lord, F. M. and Novick, M. R. Statistical theories of mental test scores. Chapter 13. The selection of predictor variables. Reading, Mass.: Addison-Wesley, 1968, 284-301.

Mills, R. G. and Williges, R. C. Performance prediction in a single-operator simulated surveillance system. Human Factors, 1972, in press.

Williges, R. C. and North, R. A. Prediction and cross-validation of video cartographic symbol location performance. Human Factors, 1972, in press.

TABLE 1

Coded Data Points of the RSM Central-Composite Design Used in the Mills and Williges (1972) Study

Treatment Condition	Blip/Scan Ratio	Target Introduction Rate	Clutter Replacement Probability	Clutter Density	Target Velocity
1	-1	-1	-1	-1	1
2	1	-1	-1	-1	-1
3	-1	1	-1	-1	-1
4	1	1	-1	-1	1
5	-1	-1	1	-1	-1
6	1	-1	1	-1	1
7	-1	1	1	-1	1
8	1	1	1	-1	-1
9	-1	-1	-1	1	-1
10	1	-1	-1	1	1
11	-1	1	-1	1	1
12	1	1	-1	1	-1
13	-1	-1	1	1	1
14	1	-1	1	1	-1
15	-1	1	1	1	-1
16	1	1	1	1	1
17	-2	0	0	0	0
18	2	0	0	0	0
19	0	-2	0	0	0
20	0	2	0	0	0
21	0	0	-2	0	0
22	0	0	2	0	0
23	0	0	0	-2	0
24	0	0	0	2	0
25	0	0	0	0	-2
26	0	0	0	0	2
27	0	0	0	0	0

TABLE 2

Additional Data Points Added to the Mills and Williges (1972) Study to Complete the 2^5 Factorial Design

Treatment Condition	Blip/Scan Ratio	Target Introduction Rate	Clutter Replacement Probability	Clutter Density	Target Velocity
1	1	1	1	1	-1
2	-1	1	1	1	1
3	1	-1	1	1	1
4	-1	-1	1	1	-1
5	1	1	-1	1	1
6	-1	1	-1	1	-1
7	1	-1	-1	1	-1
8	-1	-1	-1	1	1
9	1	1	1	-1	1
10	-1	1	1	-1	-1
11	1	-1	1	-1	-1
12	-1	-1	1	-1	1
13	1	1	-1	-1	-1
14	-1	1	-1	-1	1
15	1	-1	-1	-1	1
16	-1	-1	-1	-1	-1

TABLE 3

Multiple Correlation Coefficients

Dependent Variable	Original R	Shrunken R	Predictive Validity
	$\hat{R}_{11}$	R_S	$\hat{R}_{12}$
Probability of Correct Initiation	.818	.771	.751
Mean Initiation Latency	.760	.693	.712

TABLE 4

Comparison of Mean Observed Probability of Correct Initiation and Initiation Latency to 95 Percent Confidence Interval of Mills and Williges (1972) Prediction Equations

Treatment Condition	Mean Observed Probability of Correct Initation	95 Percent Confidence Interval of Prediction Equation	Mean Observed Initiation Latency	95 Percent Confidence Interval of Prediction Equation
1	0.75	0.69 ± .09	162.31	210.62 ± 29.14
2	0.36	0.34 ± .09	204.33	225.75 ± 29.14
3	0.93	0.87 ± .09	103.68	117.38 ± 29.14
4	0.54	0.46 ± .09	229.83	272.65 ± 29.14
5	0.77	0.78 ± .09	141.02	179.96 ± 29.14
6	0.47	0.45 ± .09	239.39	259.71 ± 29.14
7	0.91	0.90 ± .09	102.88	137.35 ± 29.14
8	0.52	0.49 ± .09	170.34	159.95 ± 29.14
9	0.81	0.85 ± .09	120.91	131.27 ± 29.14
10	0.42	0.45 ± .09	247.96	278.01 ± 29.14
11	0.88	0.91 ± .09	137.57	136.00 ± 29.14
12	0.52	0.46 ± .09	143.37	192.19 ± 29.14
13	0.80	0.79 ± .09	142.44	170.17 ± 29.14
14	0.49	0.53 ± .09	184.98	167.44 ± 29.14
15	0.94	1.00 ± .09	89.40	79.50 ± 29.14
16	0.63	0.62 ± .09	204.22	228.50 ± 29.14

TABLE 5

Summary of Significant F Ratios Resulting from Analysis of Variance of Probability of Correct Initiation and Mean Initiation Latency

		Dependent Variable	
Effect	df	Probability of Correct Initiation	Mean Initiation Latency
Blip/Scan Ratio (BSR)	1, 3	45.26**	35.94**
Target Introduction Rate (TIR)	1, 3	NS[a]	31.72*
Clutter Replacement Probability (CRP)	1, 3	93.12**	35.76**
Clutter Density (CD)	1, 3	25.16*	NS
Target Velocity (TV)	1, 3	NS	30.83*
BSR x CRP	1, 3	19.55*	NS
BSR x TV	1, 3	18.90*	13.42*
TIR x CRP x TV	1, 3	NS	11.40*
TIR x CD x TV	1, 3	NS	25.36*
CRP x CD x TV	1, 3	12.28*	NS
BSR x TIR x CD x TV	1, 3	NS	13.50*
BSR x CRP x CD x TV	1, 3	NS	10.15*
TIR x CRP x CD x TV	1, 3	NS	11.96*

[a] NS nonsignificant, $p > .05$

* $p < .05$

** $p < .01$

BIOGRAPHY

ROBERT C. WILLIGES (Predictive Validity of Central-Composite Design Regression Equations) is the Assistant Head for Research of the Aviation Research Laboratory of the Institute of Aviation and Associate Professor of Psychology and of Aviation at the University of Illinois. He received his A.B. degree in psychology from Wittenberg University in 1964 and his M.A. and Ph.D. degrees in engineering psychology from The Ohio State University in 1966 and 1968, respectively. While at Ohio State he was a research assistant at the Human Performance Center and conducted research on team training and monitoring of complex computer-generated displays. Prior to joining the Aviation Research Laboratory in 1970, he was Assistant Director of the Highway Traffic Safety Center at the University of Illinois. His current research interests include problems of visual monitoring performance, inspector behavior, and human performance in complex system operation including investigation of rate-field, frequency-separated, and visual time-compressed displays, interpretability of TV-displayed cartographic information, transfer of training, and applications of response surface methodology.

BIOGRAPHY

ROBERT G. MILLS (Predictive Validity of Central-Composite Design Regression Equations) is an engineering psychologist in the Systems Effectiveness Branch, Human Engineering Division of the Aerospace Medical Research Laboratory, Wright-Patterson AFB, Ohio. He received his B.A. and M.A. degrees from Southern Illinois University in 1961 and 1964, respectively. He is presently working toward a Ph.D. degree in industrial engineering from The Ohio State University under government sponsorship. His areas of interest include man-machine system effectiveness, computer simulation and modeling, and decision making.